利用字形风格约束的字符识别研究

刘长松 [1,2] 丁晓青 [1,2]

摘 要 印刷体字符的字形风格和手写字符的书写风格是非常重要的特性. 本文研究了利用字符字形风格之间的约束关系提高识别率的理论和方法, 提出了以字形风格同现概率为基础的 3 种识别模型, 结合实验结果分析了这些模型的优缺点和适用条件, 结果验证了本文提出的风格约束模型能够有效地提高识别率.

关键词 字符识别, 风格, 同现概率, 模板匹配
中图分类号 TP18

Study of Character Recognition Using Writing Style Consistent

LIU Chang-Song[1] DING Xiao-Qing[1]

Abstract Typeface style and writing style are important features of printed and handwritten characters, but they are not thoroughly studied for character recognition. In this paper, we study how to improve character recognition accuracy by using style consistence between characters. We propose three methods which are based on the co-occurrence probability of styles, and we analyze these methods′ features and their suitable conditions with experiments. The result shows that our style consistent models can improve recognition accuracy effectively.

Key words Character recognition, style, co-occurrence probability, template matching

1 引言

一般的字符识别研究中, 人们最关心的是输入模式所属的字符类别的问题, 并对此进行了深入的研究. 但是很多情况下, 同一类别的每个字符图像除了具有类别信息外, 还有一些个性特点, 例如, 印刷体字符具有不同字体, 手写汉字具有楷书、行书、草书等不同风格的写法. 这些个性特点一般称为风格 (Style), 由于风格的表现形式各式各样, 本文把与单个字符图像变化有关的所有风格统称为字形风格, 简称为风格.

对于印刷体字符, 字体是最明显的风格, 而且风格不仅决定于字体变化, 甚至跟环境有关, 如打印浓淡、背景噪声大小都可以认为是不同的风格. 因此, 风格不容易严格定义和确定, 是一个相对模糊的概念.

样本的风格变化常常会影响样本分类的准确性. 由于风格的复杂性, 人们对样本中出现了多少种风格变化、风格对识别的影响、以及风格间关系的研究较少. 一般采用通过对覆盖了所有类别和风格变化的大量样本学习来适应不同风格的变化, 减少样本风格变化对样本分类的准确性的影响.

文字不仅具有不同的风格, 而且连续出现的各个字的风格之间具有很强的约束关系. 因为文字一般是成串出现的, 一般情况下成串出现的文字的风格具有相似特点. 例如, 对于印刷文字, 往往某个整段落都是用同种字体和字号; 每个人的手写文字都有其独特的风格. 本文把具有连贯风格的一串文字字形风格间的关系定义为 "字形风格连续性约束". 风格连续性约束提供了字符类别以外的信息, 如何描述和利用这种信息提高字符串的识别率是本文的研究目标.

如何利用风格约束信息提高字符串的识别率, 比较直观的思路是先确定整串输入样本的字形风格, 然后使用为特定字形风格设计的分类器来进行整串输入样本识别. 在印刷体识别领域这种思路是普遍存在的, 许多字体识别方面的研究都有这种目的[1~4]. 但这种方案有两大问题: 一是要求有识别率极高的字形风格识别, 如果字形识别错误会极大降低识别率, 而识别字形风格本身是一个比识别类别更困难的问题; 二是遇到设计分类器时未建模的风格的样本时, 特定风格的分类器鲁棒性远不如风格无关的分类器.

文献 [5] 提出了 Style Consistent (即本文中的风格连续性) 的概念. 它为每个类别的每个字形风格建立在该风格条件下的混合高斯模型, 用 EM (Expectation-Maximization) 算法估计模型参数. 识别时考虑字形风格连续性, 在一定简化条件下对多个连续风格的待识别样本联合求取最优解. 这种

收稿日期 2006-4-3 收修改稿日期 2006-6-27
Received April 3, 2006; in revised form June 27, 2006
国家自然科学基金 (60472002) 资助
Supported by National Natural Science Foundation of China (60472002)
1. 清华大学智能技术与系统国家重点实验室 北京 100084 2. 清华大学电子工程系 北京 100084
1. State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084 2. Department of Electronics Engineering, Tsinghua University, Beijing 100084
DOI: 10.1360/aas-007-1121

方法能够利用风格连续性约束提高识别率, 但模型复杂, 计算量大, 且不能利用已经存在的成熟识别算法.

一些文献研究如何区分各种风格, 如文献 [6] 试图利用大量未知风格的样本自动确定增值税发票上打印字符的典型风格情况, 文献 [7, 8] 研究了联机手写英文字符的字形风格聚类.

本文提出了一套利用同现概率描述风格连续性约束的理论和方法, 并通过实验验证了利用这种约束对提高分类器的识别率的作用, 错误率降低了 10% 左右.

2 字形风格约束的描述模型和算法

2.1 具有连续性约束的字形风格模型

本文认为字形风格连续性约束可以分成两种类型:

1) 整体风格约束

每种风格由整个字符集范围内的所有字符共同拥有. 可以为每一种风格编序号. 如果有 N 种整体风格, 则每个字都有 N 种风格, 在存在整体风格约束时, 不同字符可以拥有相同的风格.

典型例子: 印刷体中的字体约束.

2) 个体风格约束

风格与每个字符的类别有关, 每个字符类别有一种或多种风格, 具有风格连续性约束的字符的风格之间有一定的联系但不一定唯一.

典型例子: 手写文字. 某一个字的写法并不能决定其他字的写法, 但每个字都有一些典型的写法, 每个字的写法会影响但不能决定其他字的写法. 例如: 对于手写汉字, 某个人的风格可能是大多数字以行书写法为主, 但某些字又可能是草书或楷书写法.

整体风格约束可以看作是个体风格约束的一种特殊情况, 整体风格约束问题也可以用个体风格约束的模型来表达.

2.2 字形风格连续性约束下的字符类别识别模型

为了利用风格间约束来帮助识别, 我们首先建立描述风格间约束情况的概率模型.

由于每个字符类别可以有一个或多个风格, 我们把每一个字符类别的每一个风格所组成的集合记为 SS. 对于 SS 集合中的任意元素 $c \in SS$ 是某一个类别的某种风格, c 具有两种属性, 即类别属性和风格属性, 分别表示为 $\mathrm{Code}(c)$ 和 $\mathrm{Style}(c)$. 假设我们已经建立了用于识别 c 的概率模型, 用该概率模型可以得到 $P(\boldsymbol{x}/c)$, $\boldsymbol{x}$ 是输入样本的特征向量.

假设 n 个风格连续样本的特征向量分别是 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$, 每个特征向量对应的可能的风格分别是 $c_1, c_2, \cdots, c_n \in SS$. 令向量 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$, 向量 $\boldsymbol{C} = \{c_1, c_2, \cdots, c_n\}$. 我们希望求

$$\boldsymbol{C}^* = \arg\max_{\boldsymbol{C}} P(\boldsymbol{C}/\boldsymbol{X}) \tag{1}$$

作为类别和风格的最优识别结果的估计. 由于 $c_1, c_2, \cdots, c_n \in SS$, 所以识别结果中同时包括类别和风格信息. 这个公式的目标是使字符的类别和风格都正确的概率最大. 字符的类别和风格都正确的概率越大, 显然字符类别识别率也越高, 如果我们只关心字符类别的识别率, 在统计识别率时忽略风格属性即可.

根据贝叶斯公式, 且由于 $P(\boldsymbol{X})$ 与 $\boldsymbol{C}$ 无关, 得到式 (2)

$$\boldsymbol{C}^* = \arg\max_{\boldsymbol{C}} \frac{P(\boldsymbol{X}, \boldsymbol{C})}{P(\boldsymbol{X})} = \arg\max_{\boldsymbol{C}} P(\boldsymbol{X}, \boldsymbol{C}) \tag{2}$$

$$P(\boldsymbol{X}, \boldsymbol{C}) = P(\boldsymbol{X}/\boldsymbol{C})\, P(\boldsymbol{C}) = P(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n/c_1, c_2, \cdots, c_n)\, P(\boldsymbol{C}) \tag{3}$$

假设 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 互相独立, 且 $\boldsymbol{x}_i$ 只与 c_i 有关, 对式 (3) 进行简化, 得到

$$P(\boldsymbol{X}, \boldsymbol{C}) = (\prod_{i=1}^{n} P(\boldsymbol{x}_i/c_1, c_2, \cdots, c_n))P(\boldsymbol{C}) = (\prod_{i=1}^{n} P(\boldsymbol{x}_i/c_i))P(\boldsymbol{C}) \tag{4}$$

两边取自然对数得到

$$\ln(P(\boldsymbol{X}, \boldsymbol{C})) = \ln(P(\boldsymbol{C})) + \sum_{i=1}^{n} \ln(P(\boldsymbol{x}_i/c_i)) \tag{5}$$

很多常用的统计分类器基于概率模型, 例如:

1) 对于高斯模型分类器

$$\ln(P(\boldsymbol{x}/c)) = \ln(\frac{1}{(2\pi)^{\frac{l}{2}} \left|\sum_c\right|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}} \sum_c^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_c))) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}} \sum_c^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_c) + \omega_c \tag{6}$$

其中 $\boldsymbol{\mu}_c$, $\sum_c$ 是对应于模型 c 的均值和协方差矩阵, ω_c 是对应于模型 c 的一个常数项. 一般我们忽略常数项, 得到马氏距离分类器.

假设协方差矩阵为对角阵, 且对角元素相等, 各个类别的协方差矩阵相同, 则马氏距离简化成欧氏

距离

$$\ln\left(P\left(\boldsymbol{x}/c\right)\right)=-\frac{1}{2\sigma^2}\left(\boldsymbol{x}-\boldsymbol{\mu}_c\right)^{\mathrm{T}}\left(\boldsymbol{x}-\boldsymbol{\mu}_c\right) \tag{7}$$

其中 σ 是常数.

2) 对于模板匹配分类器

假设风格 c 对应的模板 t 和待匹配样本尺寸相同, 总像素数为 N, 模板图像的像素 0 变 1 的概率为 p_{01}, 1 变 0 的概率为 p_{10}, 0 变 1 的像素数为 $n1$, 1 变 0 的像素数为 $n2$, 0 变 0 的像素数为 $n3$, 1 变 1 的像素数为 $n4$. 一般来说 p_{01} 与 p_{10} 均很小, 则

$$p\left(\boldsymbol{x}/c\right)=p_{01}^{n1}\cdot p_{10}^{n2}\cdot\left(1-p_{01}\right)^{n3}\cdot\left(1-p_{10}\right)^{n4}\approx p_{01}^{n1}\cdot p_{10}^{n2} \tag{8}$$

$$\ln\left(p\left(\boldsymbol{x}/c\right)\right)\approx n1\cdot\ln\left(p_{01}\right)+n2\cdot\ln\left(p_{10}\right) \tag{9}$$

假设 $p_{10}=p_{01}$, 则

$$\ln\left(p\left(\boldsymbol{x}/c\right)\right)\approx mismatch(t,\boldsymbol{x})\cdot a \tag{10}$$

其中, $a=\ln\left(p_{01}\right)$ 是一个常数, $mismatch(t,\boldsymbol{x})$ 表示模板 t 与样本 $\boldsymbol{x}$ 间不匹配的像素点数.

得到 $P\left(\boldsymbol{x}_i/c_i\right)$ 后, 关键问题是 $P(\boldsymbol{C})$ 的求解.

假设 $c_1, c_2, \cdots, c_n$ 互相独立, 以上公式退化成一般的单字识别公式.

本文研究的风格连续约束正是体现在 $c_1, c_2, \cdots, c_n$ 的关系上, 因此, 不能假设 $P(\boldsymbol{C})$ 项中 $c_1, c_2, \cdots, c_n$ 互相独立. 从前面的定义可知 $c_1, c_2, \cdots, c_n$ 不仅有类别属性还有风格属性. 如果 $c_1, c_2, \cdots, c_n$ 只是字符类别属性, 不包含风格属性, 则 $P(\boldsymbol{C})$ 相当于语言模型, 前后类别出现的顺序是有语言信息约束的.

研究字形风格约束时, 我们暂不考虑语言信息的约束. 因此假设 $c_1, c_2, \cdots, c_n$ 只与字形风格因素有关, 与顺序无关, 在这种前提下推导 $P(\boldsymbol{C})$ 形式.

由于 $c_1, c_2, \cdots, c_n$ 的组合可以很多, 预先统计出任意组合下的概率几乎不可能. 即使能够做到, 由于运算量和存储量的问题也无法在算法中使用. 因此, 必须在一些简化的条件下进行讨论.

a) 整体风格约束方法

$P(\boldsymbol{C})$ 在 $c_1, c_2, \cdots, c_n$ 风格一致时有值, 风格不一致时为零. 假设所有风格出现的频率相同, 则

$$P\left(c_1,c_2,\cdots,c_n\right)=\begin{cases}a & c_1,c_2,\cdots,c_n \quad \text{风格全部相同}\\ 0 & \text{其他}\end{cases} \tag{11}$$

其中 a 为常数. 这种假设实现简单, 但只在理想情况下成立. 相应的识别公式为

$$\boldsymbol{C}^*=\underset{c_1,c_2,\cdots,c_n\text{风格全部相同}}{\arg\max}\prod_{i=1}^{n}P(\boldsymbol{x}_i/c_i) \tag{12}$$

对于最近邻分类器的具体计算方法是把所有相同风格的模板分成一组, 分别用每种风格的模板组识别并记录该组模板识别的距离之和, 挑选距离之和最小的一组模板对应的识别结果作为最终结果.

b) 个体风格约束方法 1

每次只考虑两个输入样本间的约束关系, 此时, $P(\boldsymbol{C})=P\left(c_1,\,c_2\right)$.

本文中称 $P\left(c_1,\,c_2\right)$ 为风格 c_1, c_2 的同现概率. 欧氏距离时的优化函数

$$g\left(c_1,\,c_2\right)=\ln\left(P\left(c_1,\,c_2\right)\right)-\frac{1}{2\sigma^2}\left(\boldsymbol{x}_1-\boldsymbol{\mu}_{c1}\right)^{\mathrm{T}}\left(\boldsymbol{x}_1-\boldsymbol{\mu}_{c1}\right)-\frac{1}{2\sigma^2}\left(\boldsymbol{x}_2-\boldsymbol{\mu}_{c2}\right)^{\mathrm{T}}\left(\boldsymbol{x}_2-\boldsymbol{\mu}_{c2}\right) \tag{13}$$

模板匹配距离时的优化函数

$$g\left(c_1,c_2\right)=\ln\left(P\left(c_1,c_2\right)\right)-mismatch\left(t_1,\boldsymbol{x}_1\right)\cdot a-mismatch\left(t_2,\,\boldsymbol{x}_2\right)\cdot a \tag{14}$$

如果预先知道每个训练样本的风格, 则 $P\left(c_1,\,c_2\right)$ 可以直接统计出来. 统计方法是: 由于字形风格与字符出现的顺序无关, 在每一组风格连续的训练样本序列中, 任取两个样本, 统计两个样本的风格分别是 c_1, c_2 出现的次数, 除以所有可能的取法的总数即可得到 $P\left(c_1,\,c_2\right)$.

在实际情况下, 一般只能知道某些样本风格连续, 例如, 同一段落或同一区域的字具有连续的风格, 但具体是什么风格并不知道. 当样本的风格不知道时, 可以先利用单字识别器进行识别, 利用识别结果类别正确的首选模板的风格估计同现概率.

一次输入两个风格连续样本时, 由于只利用前后两个字间比较少的风格约束信息, 识别率提高余地不大. 而一般情况下一次识别的风格连续样本都多于两个, 针对这种情况本文提出一种解决方案.

当识别 N 个的风格连续样本时, 先把所有的风格连续样本两两组合, 分别用本文中的一次识别两个样本的方法识别, 每次组合识别都会为每个字得到一个识别结果, 这样每个字都得到 $N-1$ 个识别结果. 再将每个字符的多个识别结果集成, 本文中采用简单的投票法来集成.

此方法由于可以利用更多字间的约束信息, 不易受个别字风格畸变的干扰. 由于对每个字分别估计风格, 而不是试图统计出整体的风格再去识别, 每个字的风格识别是否正确对其他字的影响较小.

c) 个体风格约束方法 2

每次考虑多个样本间的约束关系. 假设某两个可能的识别结果 $\boldsymbol{C}$ 与 $\boldsymbol{C}'$ 只有第一个元素不

同, 不妨假设第一个元素分别为 c_1 和 c_1', 其他元素 $c_2, \cdots, c_n$ 与 $c_2', \cdots, c_n'$ 相同. 则

$$\frac{P(\boldsymbol{X}, \boldsymbol{C})}{P(\boldsymbol{X}, \boldsymbol{C}')} = \frac{(\prod_{i=1}^{n} P(\boldsymbol{x}_i/c_i))P(\boldsymbol{C})}{(\prod_{i=1}^{n} P(\boldsymbol{x}_i/c_i'))P(\boldsymbol{C}')} =$$

$$\frac{P(\boldsymbol{x}_1/c_1)P(\boldsymbol{C})}{P(\boldsymbol{x}_1/c_1')P(\boldsymbol{C}')} = \frac{P(\boldsymbol{x}_1/c_1)P(c_1/c_2, \cdots, c_n)}{P(\boldsymbol{x}_1/c_1')P(c_1'/c_2, \cdots, c_n)} \tag{15}$$

只要得到 $P(c_1/c_2, \cdots, c_n)$ 即可求解, 但是这个公式理论上还是无法求解. 本文用以下近似来估计它

$$\begin{aligned} P(c_1/c_2, \cdots, c_n) &= \frac{P(c_1, c_2, \cdots, c_n)}{P(c_2, \cdots, c_n)} = \\ &\frac{P(c_2, \cdots, c_n/c_1)P(c_1)}{P(c_2, \cdots, c_n)} = \\ &\frac{P(c_2/c_1)\cdots P(c_n/c_1)}{P(c_2)\cdots P(c_n)}P(c_1) = \\ &\frac{P(c_2, c_1)\cdots P(c_n, c_1)}{P(c_2)\cdots P(c_n)}P(c_1)^{-(n-2)} \end{aligned} \tag{16}$$

这里我们只考虑 c_1 与其他字间的关系, 假设 $c_2, \cdots, c_n$ 之间互相独立. 式 (16) 并不一定是最好的估计, 如果估计得更准确, 应该可以取得更好的结果.

通过式 (16), 我们可以分别对第一个字的第 m 个识别候选字计算一个修正的分数

$$\begin{aligned} \lambda_m &= \ln\left(P(\boldsymbol{x}_1/c_1^m) \cdot P(c_1^m)^{-(n-2)} \prod_{i=2}^{n} P(c_i, c_1^m)\right) \\ &= \ln\left(P(\boldsymbol{x}_1/c_1^m)\right) - (n-2) \cdot \ln P(c_1^m) + \\ &\quad \sum_{i=2}^{n} \ln P(c_i, c_1^m) \end{aligned} \tag{17}$$

其中 c_1^m 表示第一个字的第 m 个候选结果.

用 λ_m 对单字识别器得到的所有识别候选字重新排序, 得到修正的识别结果.

3 在字形风格连续性约束下的字符识别实验结果

3.1 已知风格的打印样本

图 1 中 6 种字体的 10 个数字, 用 6 磅字号打印, 再用 200 dpi 分辨率扫描得到灰度图像, 打印和扫描参数采样过程中保持稳定. 每个字体的每个字符有 500 个样本, 前 250 组作训练集, 后 250 组作测试集[5].

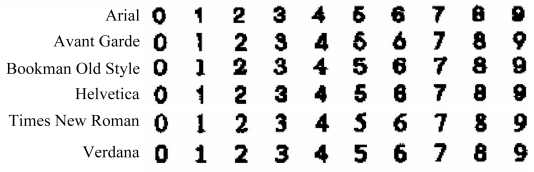


图 1 已知风格的样本

Fig. 1 Known style samples

识别特征: 对输入字的图像在 32×32 的空间中分别水平投影和垂直投影, 得到 $32 + 32 = 64$ 维特征.

由于每个样本的风格已知, 因此可以训练出每个字符每种风格的识别模型, 本文采用欧氏距离分类器模型.

如果不考虑风格, 只为每个类别训练一个模板, 得到多字体混合训练的单模板分类器识别率为 84.75%. 用每一种字体训练的模板分别识别所有字体的结果如表 1 (见下页) 所示, 从中可以看出, 测试和训练字体相同时得到最高的识别率, 也就是单字体识别器的平均识别率为 96.76%.

如果同时使用所有单字体识别器的模板来识别所有测试样本, 得到的平均识别率为 96.06%.

使用本文提出的风格约束方法得到的识别结果如表 2 (见下页) 所示. 当连续风格字符个数等于 1 时实际上等于不使用风格约束的单字识别结果.

由实验数据可知, 在比较理想的情况 (有足够多带风格信息的样本) 下, 字形风格约束分类器能够逼近单字体识别器的水平. 但如果风格连续样本长度过短, 识别率也可能会降低.

三种方案都取得了明显的识别率提高效果, 而个体约束 1 方法表现最好.

3.2 增值税发票密文识别

用文献 [6] 中的方法及相同的训练集和测试集, 分级聚类得到 27 个模板, 不断删除作用最小的模板, 到剩余 14 个模板后, 用模板匹配分类器进行识别. 得到的结果如表 3 (见下页) 所示.

由表 3 的结果可知, 整体约束导致识别率下降, 原因是从训练集得到的模板与测试集的情况具有较大的不一致性.

对于实际情况, 个体约束方法 1 具有很好的鲁棒性, 能够明显提高识别率.

个体约束方法 2 未取得好的结果, 而且在约束字符数增加的情况下, 反而识别率快速下降, 说明我们的模型在这种条件下有一定不足, 本文将在后面具体分析产生这一情况的原因.

表 1 分别用单字体训练的识别器识别结果

Table 1 Recognition results of recognizers trained by single font samples

	训练字体					
测试字体	Arial	Avantg	Bookos	Helvet	Roman	Verdan
Arial	93.24	54.12	62.06	91.88	56.56	69.28
Avantg	66.04	99.32	49.76	63.72	76	87.28
Bookos	55.32	38.92	96.96	50.36	42.2	56.28
Helvet	93.16	55.96	59.32	93.12	55.72	67.92
Roman	67.92	79.24	35.4	61.72	99.44	73.12
Verdan	91.4	67.08	53.2	85.36	65	98.48
平均	77.85	65.77	59.54	74.36	65.82	75.39

表 2 已知风格打印样本的风格约束识别结果

Table 2 Style consistent recognition results for known style printed samples

连续风格字符个数	1	2	3	4	5	6	7	8	9	10
整体约束		95.91	96.38	96.68	96.67	96.75	96.75	96.73	96.69	96.73
个体约束 1	96.06	96.51	96.59	96.85	96.77	96.91	96.91	96.96	96.93	96.92
个体约束 2		96.41	96.83	96.73	96.67	96.73	96.71	96.68	96.64	96.54

表 3 增值税发票的风格约束识别实验结果

Table 3 Style consistent recognition results for tax form

风格约束字符个数	1	2	3	4	6	12	21	42	84
整体约束		99.06	98.94	98.90	98.83	98.84	98.85	98.83	98.83
个体约束 1	99.20	99.24	99.27	99.28	99.30	99.27	99.25	99.28	99.28
个体约束 2		99.20	99.21	99.20	99.21	99.19	99.09	98.92	98.67

表 4 零概率对已知风格打印样本的识别结果影响

Table 4 Effects of zero probability on known style printed samples

相同风格字符个数	2	3	4	5	6	7	8	9	10
个体约束 2	96.41	96.83	96.73	96.67	96.73	96.71	96.68	96.64	96.54
不处理 0 概率	96.21	96.60	96.67	96.57	96.70	96.61	96.57	96.54	96.44

表 5 增值税发票密文的实验结果的改进

Table 5 Improved style consistent recognition results for tax form

风格约束字符个数	2	3	4	6	12	21	42	84
不处理零概率	99.20	99.19	99.17	99.14	99.14	98.82	98.42	97.64
个体约束 2	99.20	99.21	99.20	99.21	99.19	99.09	98.92	98.67
半数约束	99.20	99.21	99.21	99.22	99.27	99.27	99.25	99.17

3.3 风格约束纠正的错误分析

为了比较直观地理解风格约束的作用原理, 我们把风格连续约束作用下的识别结果与简单的单字识别结果进行比较, 对两者结果不同的情况进行直观的分析. 图 2、图 3 中以三个字符为一组的图像内容为: 第一个字符为待识别样本, 第二个为不考虑风格约束时最佳匹配的模板图像, 第三个为经过风格连续约束修正后最佳匹配的模板图像.

风格约束使结果正确的典型情况如图 2 所示. 从中可看出, 输入样本由于干扰与错误的模板匹配最好, 但通过风格连续性约束, 最终找到了与之风格相近的正确模板.

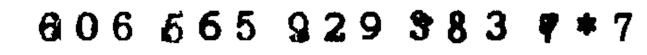
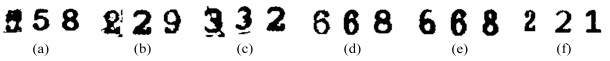


图 2 风格约束使识别结果正确的例子

Fig. 2 Style consistent recognition gives correct result

风格约束使结果错误的情况分成两种典型的情况. 第一种如图 3(a)∼(c) 所示, 即使能通过风格约束找到正确的模板, 但由于干扰太大, 仍然不能识别正确, 而单纯的模板匹配刚好能够与一个虽然风格不同但匹配度更高的模板匹配; 第二种情况如图 3(d)∼(f) 所示, 在所有的识别模板中不存在与输入样本风格一致的模板, 风格约束只会起副作用.



(a) (b) (c) (d) (e) (f)

图 3 风格约束使识别结果错误的例子

Fig. 3 Style consistent recognition gives wrong result

4 讨论

从实验结果来看, 字形风格约束识别能够很好地改进识别率, 但也出现了一些奇怪的现象, 尤其是个体约束方法 2. 下面对各种问题产生的原因进行分析.

1) 同现概率估计偏差

由于一般情况下训练样本中没有风格信息的标注, 我们只有用识别结果估计风格来统计同现概率, 这会产生一些偏差. 当识别器的识别率比较高的时候误差较小. 如果识别率较差, 则无法准确识别出风格并统计出风格间的关系, 而且由于这种情况下的主要问题已经不是由风格变化造成的, 不适合于应用风格约束的分类器, 难以用风格约束信息提高识别率.

统计同现概率需要大量的训练样本, 当训练样本不足时, 就会出现统计不充分的情况, 从而影响可靠性. 经常会发生某些风格组合的同现概率为零的情况, 零概率对概率模型具有很坏的影响, 前面的实验中, 我们都用一个很小的概率代替零概率, 这样做对个体约束方法 1 几乎没有任何效果, 对个体约束方法 2 的效果如表 4 (见上页) 所示.

小概率值对个体约束方法 2 的影响很大. 而由于统计样本稀疏, 越小的同现概率值统计误差越大, 为了进一步测试这一问题的影响, 把个体约束方法 2 中风格连续的每个字对当前处理字的同现概率排序, 只取对当前字同现概率最大的一半字符参与计算当前字的修正分数, 得到的结果如表 5 (见上页) 所示.

由以上实验可见, 个体约束方法 2 对同现概率的准确性依赖较高, 经过一定技术处理, 该方法也能取得不错的效果, 但仍不如个体约束方法 1 鲁棒性高.

2) 遇到风格没有被建模的字符样本

这种情况在实际系统中是很常见的, 这时候由于风格没有被正确建模, 利用越多的上下文字符风格信息, 副作用越大. 而且, 对整体约束和个体约束方法 2 的影响要远远大于个体约束方法 1. 这是在实用系统中, 个体约束方法 1 表现比较突出的主要原因.

3) 单字识别模型不准确

本文的计算公式都是建立在单字识别概率模型准确的基础之上, 如果其概率模型不准确将导致风格连续性约束模型不能取得最优参数.

本文的实验都是基于印刷体字符识别的, 但本文给出的模型实际上是手写或印刷体都适用的. 由于手写体文字之间的风格约束要弱于印刷体, 虽然也能通过风格约束对识别率有所提高, 但效果不如印刷体明显. 由于篇幅的原因, 本文没有给出手写体的实验结果.

5 结论

利用书写风格连续性约束确实可以明显提高识别率, 两个字不同风格之间的同现概率对于风格约束的描述至关重要, 值得进一步分析和利用.

本文方法只能改善由风格因素造成的错误, 对于其他因素的错误没有作用.

本文提出了整体约束方法、个体约束方法 1、个体约束方法 2, 共三种模型, 其中个体约束方法 1 性能最好, 且最容易使用. 该方法不需要人工标定风格属性的训练样本, 易于使用; 也不需要特殊形式的分类器模型, 容易用于优化许多现有系统.

本文的模型要求每个字的每种风格有一个独立的模板. 本文提出的方法不仅能够识别出类别, 还能够识别出风格信息.

存在问题: 如果用风格约束方法识别未建模的风格的样本性能可能更坏; 存储同现概率会增加额外的消耗, 存储量与风格个数的平方成正比; 对于大

字符集识别问题的可用性受到限制; 需要较多的训练样本来得到风格同现概率和每种风格的识别模板.

References

1 Öztürk S, Sankur B, Abak T. Font classification and recognition in document images. *Journal of Electronic Imaging*, 2001, **10**(2): 418∼430

2 Lee C W, Kang H, Jung K, Kim H J. Font classification using NMF. In: Proceedings of the 10th International Conference on Computer Analysis of Images and Patterns. Groningen, Netherlands: Springer, 2003. 470∼477

3 Shi H, Pavlidis T. Font recognition and contextual processing for more accurate text recognition. In: Proceedings of 4th International Conference on Document Analysis and Recognition. Ulm, Germany: IEEE Press, 1997. 39∼44

4 Zramdini A, Ingold R. Optical font recognition using typographical features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(8): 877∼882

5 Prateek S, George N. Style consistent classification of isogenous patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(1): 1∼11

6 Liu C S, Ding X Q. Automatic style clustering of printed characters in form images. In: Proceedings of SPIE. San Jose, USA: SPIE, 2005. 175∼182

7 Vuurpijl L G, Schomaker L R. Coarse writing-style clustering based on simple stroke-related features. In: Proceedings of the International Workshop on Fronties of in Handwriting Recognition. London: World Scientific, 1997. 29∼34

8 Vuokko V. Clustering writing styles with a self-organizing map. In: Proceedings of the 8th International Workshop on Fronties of in Handwriting Recognition. Ontario, Canada: IEEE Press, 2002. 345∼350

刘长松 清华大学电子工程系副教授, 主要研究方向为文本图像处理、模式识别、自然语言处理. 本文通信作者.
E-mail: lcs@mail.tsinghua.edu.cn
(**LIU Chang-Song** Associate professor in Department of Electronics Engineering at Tsinghua University. His research interest covers document image processing, pattern recognition, and natural language processing. Corresponding author of this paper.)

丁晓青 清华大学电子工程系教授, 主要研究方向为图像处理、模式识别、生物特征识别、视频监控.
(**DING Xiao-Qing** Professor in Department of Electronics Engineering at Tsinghua University. Her research interest covers image processing, pattern recognition, biometric identification and authentication, and video surveillance.)