一种基于微结构特征的多文种文本无关笔迹鉴别方法

李 昕 [1, 2, 3] 丁晓青 [1, 2, 3] 彭良瑞 [1, 2, 3]

摘 要 与字符识别一样, 计算机自动笔迹鉴别是一个涉及到不同文种的研究课题. 本文提出了一种基于网格窗口微结构特征的文本无关的笔迹鉴别方法, 能适用于各种不同文种的笔迹. 该方法对笔迹中局部细微结构的书写变化趋势进行描述, 并采用加权距离度量方法进行笔迹相似性度量. 利用该方法实现了文本无关的多文种笔迹检索系统, 并在实际汉字、英文、藏文和维吾尔文的笔迹库上进行了测试. 实验证明, 该方法是一种高效且适用性较广、限制性较少的笔迹鉴别方法.

关键词 笔迹鉴别, 文本无关, 多文种, 微结构特征, 加权距离度量
中图分类号 TP391.4

A Microstructure Feature Based Text-independent Method of Writer Identification for Multilingual Handwritings

LI Xin[1, 2, 3] DING Xiao-Qing[1, 2, 3] PENG Liang-Rui[1, 2, 3]

Abstract As the same as character recognition, computer automatic writer identification is a research subject involving different languages. This paper proposes a text-independent method of writer identification based on grid-window microstructure feature for different multilingual handwritings. The proposed method depicts the writing trend of local fine structures in handwritings and uses weighted distance metrics to measure the similarity between handwritings. Based on the proposed method, a text-independent handwriting retrieval system is implemented. The system is tested on the real handwriting databases of Chinese handwriting, English handwriting, Tibetan handwriting, and Uighur handwriting. The experimental results demonstrate that the proposed method is a general-purpose and weak-confined method of writer identification with high efficiency.

Key words Writer identification, text-independent, multilingual, microstructure feature, weighted distance metric

近几年来, 计算机自动笔迹鉴别成为在国际上相当活跃的一个研究领域. 与文字识别技术一样, 笔迹鉴别也是一个涉及不同文种的问题. 国内外有不少机构致力于研究各种不同文种笔迹的自动识别技术. 而现阶段, 适用于多种文种的通用的计算机自动笔迹鉴别技术还不成熟.

现有的计算机笔迹鉴别方法主要分为文本相关方法、文本无关方法[1] 和半文本无关方法[2]. 文本相关和半文本无关方法需要通过手写字符识别或人工标定获得文本的内容信息, 即需要了解文本中含有哪些字符. 文本无关方法则不考虑文本内容, 直接从笔迹图像中提取笔迹风格信息.

收稿日期 2008-08-29 收修改稿日期 2009-01-12
Received August 29, 2008; in revised form January 12, 2009
国家重点基础研究发展计划 (973 计划) (2007CB311004), 国家自然科学基金 (60772049, 60872086) 资助
Supported by National Basic Research Program of China (973 Program) (2007CB311004) and National Natural Science Foundation of China (60772049, 60872086)
1. 清华大学智能技术与系统国家重点实验室 北京 100084 2. 清华信息科学与技术国家实验室 北京 100084 3. 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. Tsinghua National Laboratory for Information Science and Technology, Beijing 100084 3. Department of Electronics Engineering, Tsinghua University, Beijing 100084
DOI: 10.3724/SP.J.1004.2009.01199

汉字和英文无疑是在世界范围内使用最广泛的文字. 在我国, 除了汉字外, 一些地区也广泛使用着蒙文、藏文、维吾尔文和朝鲜文等民族文字. 本文选取藏文和维吾尔文两种文字的笔迹, 与汉字、英文笔迹一起作为我们的研究对象. 汉字是典型的方块结构字符, 英文则是典型的拼音文字. 汉字笔迹中的字符之间一般有明显间隔, 字符间的连写较英文要少得多. 由于汉字字符类别多、组成笔画复杂、形态差异很大, 字符拼接而成的笔迹纹理文本块的纹理变化比西方文种笔迹要大. 因此, 基于纹理分析的笔迹鉴别方法在汉字笔迹上的效率往往不如在英文笔迹上的效率高. 而文本相关的方法对汉字、日文等东方文字笔迹往往更有效[3−4]. 藏语属汉藏语系藏缅语支, 藏文是以古印度字母为基础的文字. 在实际手写藏文笔迹中, 藏文字根间往往有大量的连接笔画, 字根本身形态的变化很大, 切分识别藏文字根非常困难. 以获得文本字符信息为前提的文本相关的笔迹鉴别方法对藏文笔迹并不合适. 而维吾尔语属阿尔泰语系突厥语族, 维吾尔文采用阿拉伯文字母, 其字符形式复杂, 字符间连写更加频繁.

由于藏文和维吾尔文的手写字符识别技术还不成熟, 人工标定代价较高, 本文不考虑以单一字、词为比较单元的文本相关及半文本无关方法, 而致力

于研究适应多种文字的文本无关的笔迹鉴别方法. 国外针对阿拉伯文笔迹鉴别的研究, 也主要集中在基于纹理分析的文本无关方法上. 基于纹理分析的文本无关笔迹鉴别方法在汉字、英文和阿拉伯文等笔迹上得到了成功应用. 文献 [5−7] 利用多通道 Gabor 特征表征笔迹纹理. 文献 [8−9] 从笔迹纹理块中提取小波特征, 并建立基于统计概率框架的模型 (广义高斯分布 (Generalized Gaussian density, GGD) 模型或隐马尔科夫树 (Hidden Markov tree, HMT) 模型) 来描述笔迹. 由于笔迹纹理块生成依赖字符或文本拼接, 而字符或文本的不同组合会使纹理本身发生较大变化, 从而影响稳定笔迹特征的提取. 因此, 本文不采用基于纹理块纹理分析的方法.

Bulacu 和 Schomaker 提出了基于一系列概率分布函数 (Probability distribution function, PDF) 特征的文本无关笔迹鉴别方法, 并在英文、阿拉伯文笔迹上得到极其成功的应用[10−12]. 其主要思想是用概率分布函数来描述书写人在书写笔迹时的习惯和趋向. 概率分布函数中每一个概率值都对应某种基本写法的出现概率. 这样, 就描述了书写者使用各种基本写法的频率, 从而反映书写者的书写习惯. 而不同基本写法可以用相连的两个笔迹边缘碎片的方向组合来描述, 也可以从大量微切分碎块中聚类学习出典型的字形微粒, 如文献 [11] 中性能最好的 Contour-Hinge 特征和 Grapheme emission 特征.

借鉴概率分布函数特征的思想, 针对多文种笔迹的特点, 为了提高鉴别算法的通用性, 本文提出了一种从边缘图像上提取的网格窗口微结构特征. 这种微结构特征与文献 [11] 中的一系列特征一样, 是一种概率密度函数特征. 它将局部窗口中成对出现的边缘像素点视为局部微结构, 并以不同局部微结构在整幅图像中的概率分布函数来表征笔迹风格特性. 此外, 我们还提出了使用加权的距离度量方法来计算概率密度函数特征向量间的距离, 弥补了简单距离度量方法的不足. 基于微结构特征的笔迹鉴别方法能够适应不同文种的笔迹. 在汉字、英文、藏文和维吾尔文笔迹上的实验结果, 也证明了本文方法对多文种笔迹的良好鉴别性能.

1 笔迹特征提取

要解决多文种笔迹鉴别的问题, 首先要适应不同文种的特点, 特别是汉字等东方文字笔迹. 文献 [11] 的 Contour-Hinge 特征是可以使用在中文笔迹的, 但汉字笔画交错, 单纯的边缘方向描述可能不够细致. 而推广 Grapheme emission 特征就存在限制. 因为汉字结构复杂, 很难找到一种鲁棒的切分方法来获取稳定的切分碎片. 汉字笔画类型的多样性也大大增加了字形微粒的数量. 本文提出的特征则借鉴了 Contour-Hinge 特征和 Grapheme emission 特征的优点. 特征提取方法主要分三个步骤: 笔迹样本预处理、图像边缘检测和微结构特征提取.

1.1 笔迹样本预处理和图像边缘检测

对含有笔迹的纸张样本进行扫描, 可以获得笔迹图像. 本文中使用的笔迹图像均是以 300 dpi 扫描获得. 为保持笔迹原始形态, 不再对图像进行其他归一化的操作. 图像的预处理包括图像二值化和非笔迹信息去除 (包括去除图形、格线等), 以获得只含有墨迹像素点的二值图像. 本文之所以没有使用彩色和灰度图像, 是因为彩色和灰度信息受书写环境 (笔、纸) 等影响较大. 这里不再对具体预处理方法进行详述. 我们认为经过预处理得到的二值图像中的所有黑像素点都是书写人的行笔墨迹.

微结构特征是在边缘图像上提取的. 图像边缘检测算法很多, 有利用求导算子 (Roberts, Prewitt 和 Sobel 算子等) 检测梯度最大值, 还有检测二阶导数的过零点, 以及利用多尺度小波进行边缘检测等. 本文采用 Sobel 算子获取图像梯度, 然后利用基于噪声的均方根 (Root mean square, RMS) 误差为指标进行估计[13] 的方法来确定边缘的判决阈值. 从而, 最终获得笔迹的边缘图像.

1.2 微结构特征提取

在图像边缘上提取笔迹鉴别的特征, 是我们进行笔迹鉴别的关键. 本文提出了一种特殊的从边缘图像上提取的特征. 首先, 对边缘图像局部区域里的细微结构进行描述, 然后在全局范围内统计局部细微结构出现概率, 并用这个概率分布来描述一个人的笔迹风格. 这个概率分布函数就是提出的网格窗口微结构特征. 具体而言, 它是利用一个 $(2n+1)\times(2n+1)$ 的网格窗口, 通过让其窗口中心遍历边缘图像的所有边缘像素点来进行特征提取的, 其中 n 是窗口大小控制参数. 在遍历过程中, 我们统计各种局部微结构的出现频率来最终获得对整幅笔迹边缘图像的特征描述. 下面先给出局部微结构的定义.

图 1 的左边是一个 9×9 的网格窗口, 即 $n=4$. 我们将除中心网格以外的其他 80 个网格位置按其距中心网格的距离分成 4 组. 第 1 组是中心网格的 8 邻域, 第 m ($m\geq2$) 组是第 $m-1$ 组网格的外邻域. 由图 1 可知, 第 m ($1\leq m\leq n$) 组中的网格有 $8m$ 个. 我们将这 80 个网格分别用固定的符号 i_m 来标记, 其中 $0\leq i\leq 8m-1$, $1\leq m\leq n$. 比如, 第 1 组网格 (中心网格的 8 邻域网格) 按逆时针顺序分别被标记为 0_1,1_1,$\cdots$,7_1. 我们定义在这个局部网格窗口中, 两个同组网格位置上同时出现边

缘像素点的情形为一种局部微结构. 比如, 如果 i_m 和 j_m 上同时出现边缘像素点, 则得到一种局部微结构, 记为 (i_m, j_m). 通过考察这个网格窗口所覆盖的局部区域内两两边缘像素点的位置, 可以获得该局部区域内包含的所有局部微结构种类.

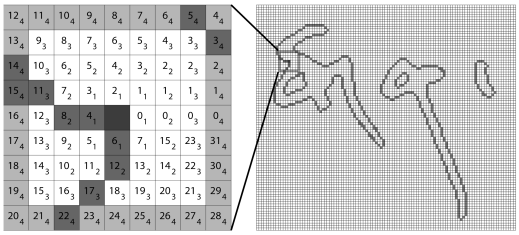


图 1　微结构特征提取示意图

Fig. 1　Schematic diagram of extracting microstructure feature

我们以一小块实际藏文笔迹的边缘图像 (如图 1 右边所示) 为例, 来说明微结构特征的提取过程. 当网格窗口的中心落在某一边缘像素点上时, 我们考察窗口中的每组网格上多个边缘像素点出现的情况. 如图 1 所示, 在第 1 组网格中, 4_1 和 6_1 网格出现边缘像素点, 而在第 4 组网格中, 3_4、5_4、14_4、15_4 和 22_4 上都出现了边缘像素点. 一般来说, 同一组网格中会至少出现两个边缘像素点. 当同一组网格中出现多个边缘像素点时, 我们不记录所有两两组合情况, 而以每两个相隔最近的边缘像素点作为一种局部微结构来记录. 如图 1 中的第 4 组网格, 我们就把这多个边缘像素点的情况看作 $(3_4, 5_4)$、$(5_4, 14_4)$、$(14_4, 15_4)$ 和 $(15_4, 22_4)$ 同时出现. 由于同组网格的两两组合情况都可能出现, 第 m 组网格可组成 C_{8m}^2 种微结构, 则总共要记录的微结构有 $\sum_{m=1}^{n} \mathrm{C}_{8m}^2$ 种. 我们用数值集合 $h(i_m, j_m)$ $(1 \le m \le n,\ 0 \le i < j \le 8m-1)$ 来记录各个微结构的出现次数.

在网格窗口遍历图像之前, 令所有 $h(i_m, j_m) = 0$. 当有某一微结构出现时, 我们就令其对应的 $h(i_m, j_m) = h(i_m, j_m) + 1$. 在图 1 所示的情况中, 需要进行自加 1 运算的 $h(i_m, j_m)$ 有: $h(4_1, 6_1)$, $h(8_2, 12_2)$, $h(11_3, 17_3)$, $h(3_4, 5_4)$, $h(5_4, 14_4)$, $h(14_4, 15_4)$ 和 $h(15_4, 22_4)$. 当网格窗口遍历完所有的边缘像素点后, 累加后的 $h(i_m, j_m)$ 就记录了各个局部微结构在整个图像中出现的次数. 最后, 我们用总次数 $\sum_{i,j,m} h(i_m, j_m)$ 对各种微结构的出现次数进行归一化, 从而得到各种微结构出现的概率密度值为

$$p(i_m, j_m) = \frac{h(i_m, j_m)}{\sum\limits_{i,j,m} h(i_m, j_m)} \tag{1}$$

这些概率密度值满足 $\sum_{i,j,m} p(i_m, j_m) = 1$. 我们就用这些概率密度值组成的概率分布函数作为描述笔迹风格的网格窗口微结构特征. 实际上, 微结构特征描述了局部区域中同组网格位置上成对出现边缘像素点的情况在全局图像上的概率分布. 微结构特征提取算法的时间复杂度主要由边缘图像中边缘像素点数目和网格窗口大小所决定. 如果以 $(2N+1) \times (2N+1)$ 的网格窗口遍历含有 M 个边缘像素点的边缘图像, 时间复杂度为 $\mathrm{O}(MN^2)$. 网格窗口的大小在很大程度影响算法的效率.

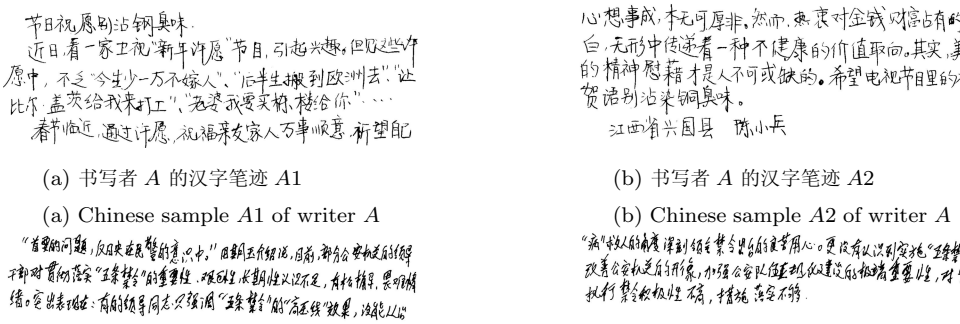
因此, 微结构特征提取时使用的窗口一般比较小, 其关注的都是局部笔画层次上的笔迹信息. 任何一种文种的笔迹都是由大量的基本笔画组成, 只是不同文种笔迹的基本笔画组成不同, 其笔画包含的局部微结构也不相同. 而在相同的文种笔迹上, 不同的书写者都有其书写的个性特点. 一般来说, 同一文种的笔画可以分为有限个类别, 比如汉字中的横、竖、撇、捺等. 对于同一类笔画, 不同人会有不同的写法, 而同一个人每次书写的写法也会有差异. 因此, 我们一般很难判断同类笔画少量样本上的差异是来自同一个书写者个人的变化, 还是来自不同书写者的风格差异. 但是, 我们提出的通过全局统计得到的微结构特征能够描述笔迹中不同细微结构的出现频率, 描述书写者更趋于使用何种微结构和不善于使用何种微结构, 这样就给出了书写者在笔画层次上的书写趋势. 比如对于某一类笔画来说, 不同人书写时会习惯使用不同的长度、角度和弧线弯曲程度. 而如果长度、角度和弯曲程度确定, 也基本确定了单个笔画的形态. 如果一个人习惯书写某类笔画比较长, 那么在更外围网格上的微结构分布就更多; 反之, 在内部网格上分布就更多. 如果一个人书写某类笔画时习惯沿某一个方向倾斜, 那么在这个方向角度上的微结构分布就会更多. 如果一个人书写该类笔画的弧线越弯曲, 那么微结构特征在方向角度上分布沿网格内外的变化就越明显; 反之, 在方向角度上分布沿网格内外的变化就更趋于一致. 另外, 笔画的粗细程度和笔画间的连写习惯在微结构特征上也有一定的体现. 这样, 书写者在书写笔画习惯上的差异, 就通过微结构特征体现出来了. 这个差异实际上就是书写人笔迹风格的差异. 与考察两个一定长度边缘碎片方向的 Contour-Hinge 特征相比, 我们提出的网格窗口微结构特征能更好地表征笔画结构的连续变化和关联属性. Contour-Hinge 特征限定了边缘碎片的长度, 仅考虑了一定距离上笔画边缘本身的延展方向. 而网格窗口微结构特征, 通过描述连续变化距离上边缘点位置, 更细致地刻画了笔画本身特性以及邻近笔画间的关系. 实际上, Grapheme emission 特征也包含了相对复杂的信息,

但微结构特征没有聚类学习的步骤. 同时, 我们选用一定大小的网格窗口, 能更鲁棒地适应同一书写人书写笔画时的细微差异, 而 Contour-Hinge 特征则对此更加敏感.

需要说明的是, 微结构特征处理的是文本无关的笔迹样本. 文本内容不同, 笔迹中各种笔画类别的出现比例就会有变化, 这同样会引起各种微结构的概率分布产生差异. 但是, 当文本中含有的字符达到一定数量后, 各种笔画类别的比例将达到统计意义上的稳定, 微结构特征也就能够主要反映书写者之间的笔迹风格差异了. 因此, 我们提出的微结构特征需要从有足够篇幅的整篇或整段文本的笔迹样本上提取.

1.3 网格窗口微结构特征示例

图 2～4 分别给出了汉字、英文和藏文笔迹的样本, 每种文种都包含两位书写者的各两篇笔迹, 文本内容各不相同. 图 5～7 则给出了从上述十二篇笔迹上提取的微结构特征示意图. 这里的微结构特征是在 $n = 4$ 条件下提取的. 由于概率值 $p(i_m, j_m)$ 满足 $1 \le m \le n$ 且 $0 \le i < j \le 8m - 1$, 图示中的概率分布集中在 4 个三角形区域内. 通过比较这 12 篇笔迹样本的微结构特征, 可以看出不同笔迹样本间的差异.



(a) 书写者 A 的汉字笔迹 $A1$

(a) Chinese sample $A1$ of writer A

(b) 书写者 A 的汉字笔迹 $A2$

(b) Chinese sample $A2$ of writer A

(c) 书写者 B 的汉字笔迹 $B1$

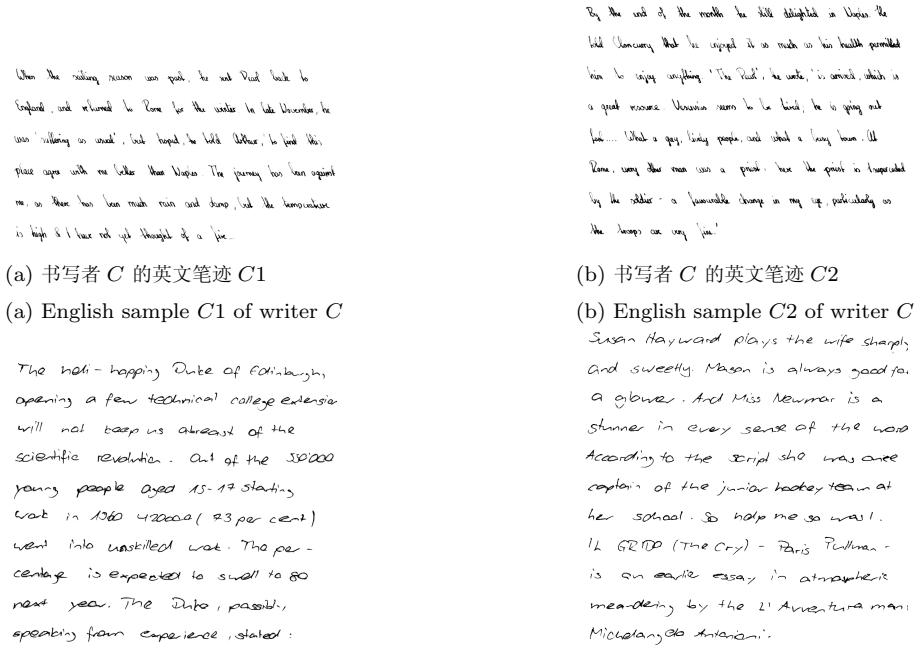
(c) Chinese sample $B1$ of writer B

(d) 书写者 B 的汉字笔迹 $B2$

(d) Chinese sample $B2$ of writer B

图 2 汉字笔迹示例

Fig. 2 Chinese handwriting samples of writers A and B



(a) 书写者 C 的英文笔迹 $C1$

(a) English sample $C1$ of writer C

(b) 书写者 C 的英文笔迹 $C2$

(b) English sample $C2$ of writer C

(c) 书写者 D 的英文笔迹 $D1$

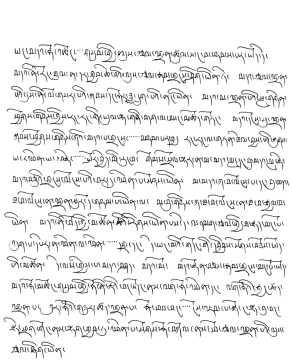
(c) English sample $D1$ of writer D

(d) 书写者 D 的英文笔迹 $D2$

(d) English sample $D2$ of writer D

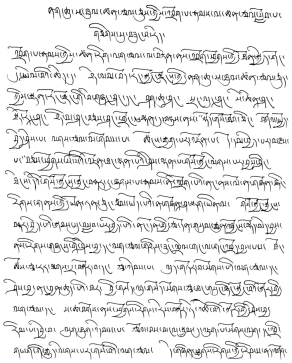
图 3 英文笔迹示例

Fig. 3 English handwriting samples of writers C and D



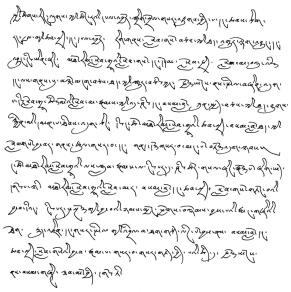
(a) 书写者 E 的藏文笔迹 E1

(a) Tibetan sample $E1$ of writer E



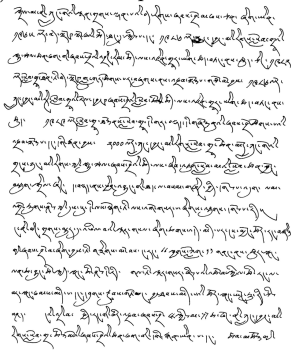
(b) 书写者 E 的藏文笔迹 $E2$

(b) Tibetan sample $E2$ of writer E



(c) 书写者 F 的藏文笔迹 $F1$

(c) Tibetan sample $F1$ of writer F

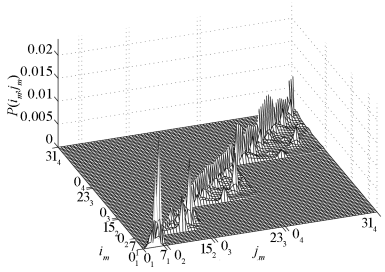


(d) 书写者 F 的藏文笔迹 $F2$

(d) Tibetan sample $F2$ of writer F

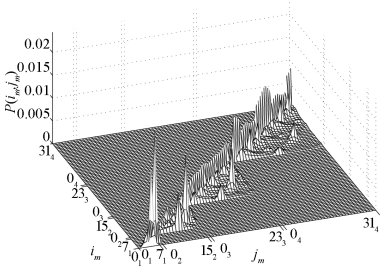
图 4　藏文笔迹示例

Fig. 4　Tibetan handwriting samples of writers E and F



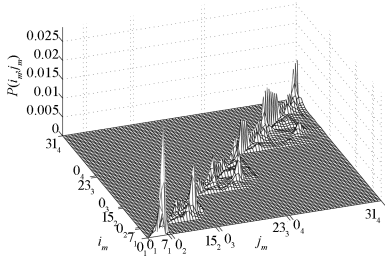
(a) 汉字笔迹 $A1$ 的微结构特征

(a) Microstructure feature of $A1$



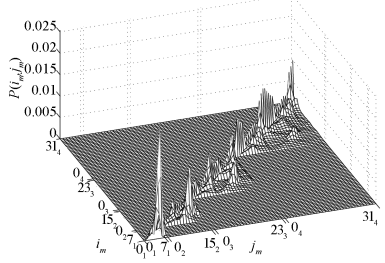
(b) 汉字笔迹 $A2$ 的微结构特征

(b) Microstructure feature of $A2$



(c) 汉字笔迹 $B1$ 的微结构特征

(c) Microstructure feature of $B1$

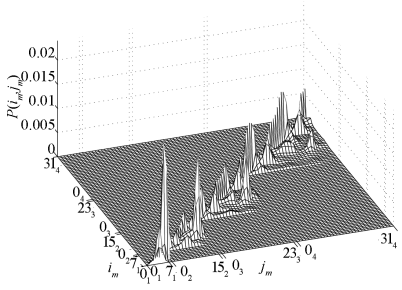


(d) 汉字笔迹 $B2$ 的微结构特征

(d) Microstructure feature of $B2$

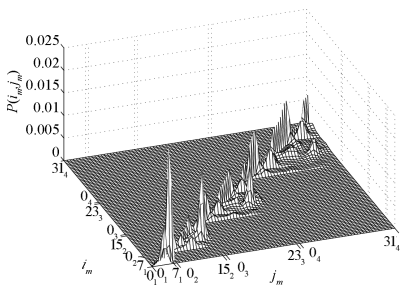
图 5　汉字笔迹的微结构特征示例

Fig. 5　Microstructure features of Chinese handwriting samples



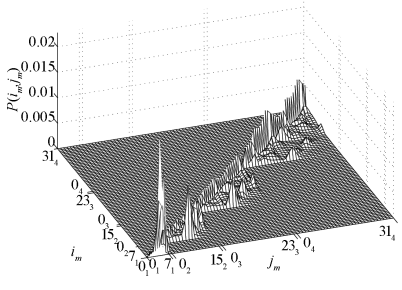
(a) 英文笔迹 $C1$ 的微结构特征

(a) Microstructure feature of $C1$



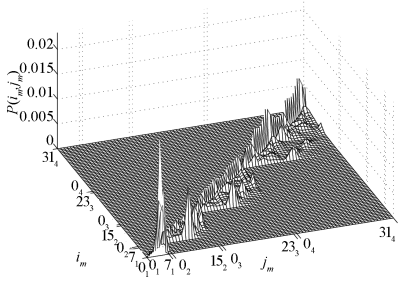
(b) 英文笔迹 $C2$ 的微结构特征

(b) Microstructure feature of $C2$



(c) 英文笔迹 $D1$ 的微结构特征

(c) Microstructure feature of $D1$

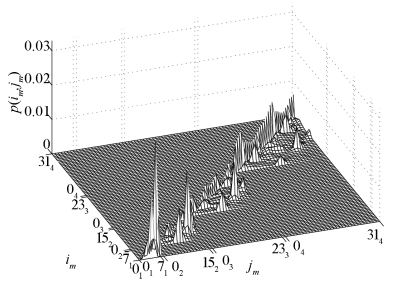


(d) 英文笔迹 $D2$ 的微结构特征

(d) Microstructure feature of $D2$

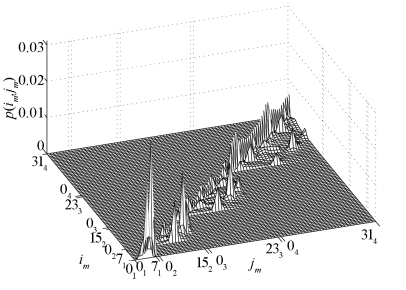
图 6 英文笔迹的微结构特征示例

Fig. 6 Microstructure features of English handwriting samples



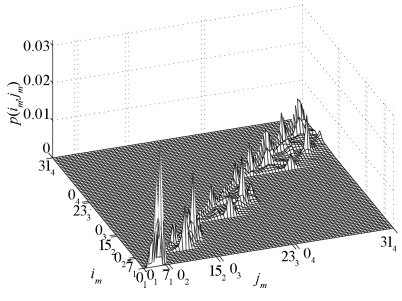
(a) 藏文笔迹 $E1$ 的微结构特征

(a) Microstructure feature of $E1$



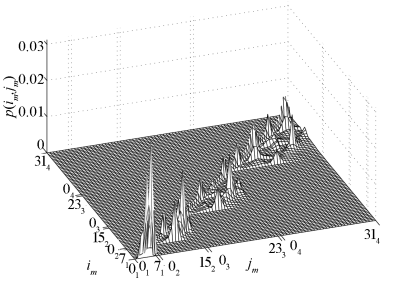
(b) 藏文笔迹 $E2$ 的微结构特征

(b) Microstructure feature of $E2$



(c) 藏文笔迹 $F1$ 的微结构特征

(c) Microstructure feature of $F1$



(d) 藏文笔迹 $F2$ 的微结构特征

(d) Microstructure feature of $F2$

图 7 藏文笔迹的微结构特征示例

Fig. 7 Microstructure features of Tibetan handwriting samples

为了更清晰地显示微结构特征间的差异, 图 8 提供了这三种文种笔迹样本间微结构特征差值的绝对值的示意图, 用来说明特征间差异在不同书写者和相同书写者条件下的情形. 由图 8 可以看出, 同一文种不同书写者笔迹的微结构特征间存在着明显的差异, 而同一书写者笔迹的微结构特征则非常相似. 因此, 我们得出结论: 同一书写者的书写习惯在统计上是稳定的, 而不同书写者的书写习惯则有所不同. 书写者笔迹风格特性的差异能够通过微结构特征反映出来. 不同文种有不同的文字特点, 其局部结构不尽相同, 但微结构特征能够反映出不同文种笔迹局部结构的呈现趋势.

2　笔迹相似性度量

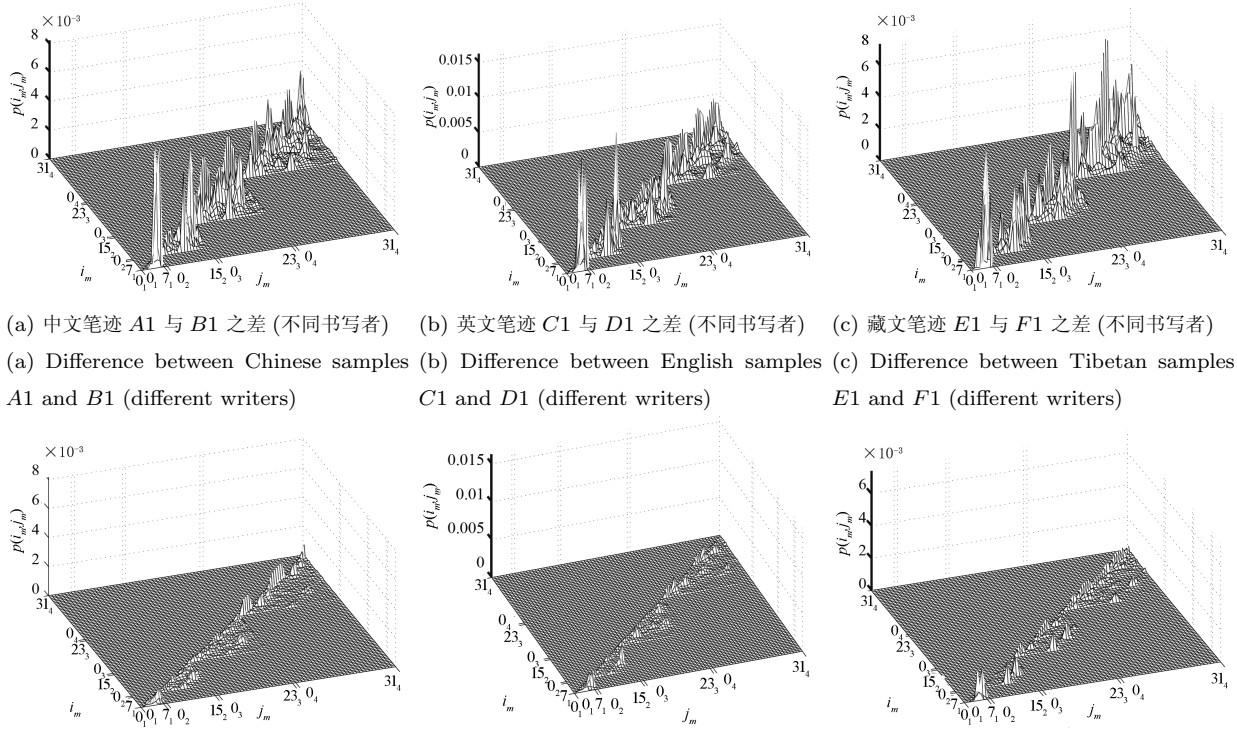
在微结构特征描述了整篇笔迹的风格特性之后, 我们需要对笔迹之间的相似性进行度量, 把笔迹之间的差异转化为可比较的数值体现. 完成这一步骤的常用方法就是距离度量, 用描述对象特性的特征向量之间的距离作为对象差异性的度量. 距离越大, 对象间差异就越大; 反之, 距离越小, 两个对象就越相似. 我们将微结构特征中的各个概率值排成一个向量. 这个向量有 $\sum_{m=1}^{n} \mathrm{C}_{8m}^{2}$ 维. 这样, 相似性度量就变成计算高维向量间的距离. 常用的距离度量方法有很多, 比如: 欧氏距离 (Euclidean distance), 巴特查里亚距离 (Bhattacharyya distance), 卡方距离 (Chi-square distance, χ^2 distance) 等. 假设有两篇笔迹的微结构特征向量为 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$, 则它们之间的欧氏距离 d_{Euc} 和卡方距离 d_{Chi} 分别为

$$d_{\mathrm{Euc}}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \sqrt{\sum_{i=1}^{N} (v_{1i} - v_{2i})^2} \tag{2}$$

和

$$d_{\mathrm{Chi}}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \sum_{i=1}^{N} \frac{(v_{1i} - v_{2i})^2}{v_{1i} + v_{2i}} \tag{3}$$

其中, v_{1i} 和 v_{2i} 分别表示 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 的各维元素, N 表示向量维数. 另外, 不同文种笔迹的笔画构成不同, 局部微结构特征也不完全相同. 有的细微结构经常出现, 有的细微结构则出现较少. 反映在微结构特征向量上, 不同维度上概率值的范围不同. 因此, 对不同维度上的数值赋予不同的权重就成为对距离度量方法的可能改进. 常用的加权方法是用不同维度上数值的标准差作为加权值. 而这一标准差可以用很多人的参考样本笔迹统计出来. 假设笔迹库中有多



(a) 中文笔迹 $A1$ 与 $B1$ 之差 (不同书写者)　(b) 英文笔迹 $C1$ 与 $D1$ 之差 (不同书写者)　(c) 藏文笔迹 $E1$ 与 $F1$ 之差 (不同书写者)

(a) Difference between Chinese samples $A1$ and $B1$ (different writers)　(b) Difference between English samples $C1$ and $D1$ (different writers)　(c) Difference between Tibetan samples $E1$ and $F1$ (different writers)

(d) 中文笔迹 $A1$ 与 $A2$ 之差 (相同书写者)　(e) 英文笔迹 $C1$ 与 $C2$ 之差 (相同书写者)　(f) 藏文笔迹 $E1$ 与 $E2$ 之差 (相同书写者)

(d) Difference between Chinese samples $A1$ and $A2$ (the same writer)　(e) Difference between English samples $C1$ and $C2$ (the same writer)　(f) Difference between Tibetan samples $E1$ and $E2$ (the same writer)

图 8　不同或相同书写者笔迹间微结构特征差异的示意图

Fig. 8　Differences of microstructure features between handwriting samples of different writers or the same writer

个参考笔迹样本, 它们的特征向量是 $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K$, 那么维度 i 上的标准差为

$$\sigma_i = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(v_{ki}-\mu_i)^2} \tag{4}$$

其中, v_{ki} 是 $\boldsymbol{v}_k$ 的第 i 个元素, 而 μ_i 是维度 i 上的均值

$$\mu_i = \frac{1}{K}\sum_{k=1}^{K} v_{ki} \tag{5}$$

用 σ_i 去除各维元素 v_{ki}, 可以得到加权向量 $\boldsymbol{v}'_k$ 的元素 $v'_{ki} = v_{ki}/\sigma_i$. 这样, 我们通过计算加权向量 $\boldsymbol{v}'_k$ 间的距离度量, 得到各种距离度量的加权版本. 加权欧氏距离和加权卡方距离分别表示为

$$d_{\mathrm{WEuc}}(\boldsymbol{v}_1, \boldsymbol{v}_2) = d_{\mathrm{Euc}}(\boldsymbol{v}'_1, \boldsymbol{v}'_2) \tag{6}$$

$$d_{\mathrm{WChi}}(\boldsymbol{v}_1, \boldsymbol{v}_2) = d_{\mathrm{Chi}}(\boldsymbol{v}'_1, \boldsymbol{v}'_2) \tag{7}$$

3 笔迹鉴别实验结果

全笔迹鉴别是从大量提供参考笔迹的候选书写人中找出查询笔迹最可能的一个或若干个书写者. 我们通过对查询笔迹和所有参考笔迹间相似度进行排序, 找到那些与查询笔迹最相似的参考笔迹的书写人, 作为查询笔迹最可能的书写者. 利用本文提出的笔迹鉴别方法, 我们实现了一个适用于多种文种的基于 Oracle 数据库的笔迹检索系统, 并在 HIT-MW 汉字手写样本库[14]、IAM 英文手写样本库[15] 以及收集的实际藏文和维吾尔文笔迹样本库上, 对基于微结构特征方法在笔迹鉴别中的性能进行了实验.

3.1 笔迹样本库

HIT-MW 汉字手写文本库和 IAM 英文手写样本库起初都是为研究手写字符识别而建立的. 我们这里利用了 HIT-MW 库中 240 人和 IAM 库中 93 人的手写文本, 这些手写文本内容各不相同. 由于 HIT-MW 库中每个书写人只提供了一篇文本, 我们人工把每篇文本分割成长度相当的两篇笔迹样本. 比如图 2 中的 $A1$ 和 $A2$ 笔迹原本就属于 HIT-MW 库中的同一篇文本, 而 $B1$ 和 $B2$ 属于另一篇文本. 实际的 IAM 库中总共有 657 人的手写文本, 我们这里选取的 93 人都提供了两篇手写文本. 笔迹样本示例如图 4. 另外, 我们自己收集了近 400 名藏族和维吾尔族书写者的笔迹, 每位书写者都提供了两篇用本族文字书写的笔迹样本. 最终, 我们得到了 266 人的藏文笔迹和 120 人的维吾尔文笔迹, 每个人都有两篇不同文本内容的笔迹样本. 藏文笔迹样本如图 6 所示. 我们的笔迹鉴别实验就在这四个笔迹库上进行.

3.2 实验过程和实验结果

我们将每个人的两篇笔迹中的一篇作为参考笔迹, 另一篇作为查询笔迹. 在汉字笔迹库中, 每篇汉字查询笔迹都与 240 篇汉字参考笔迹进行比较, 获得它们间的相似性度量, 即计算所提取微结构特征向量间的距离. 通过对距离值进行排序, 可以得到与某篇查询笔迹最相似的参考笔迹, 这篇参考笔迹的书写人就被认为是该查询笔迹最可能的书写者. 进一步, 还可以获得该查询笔迹最可能的前若干个书写人的候选名单. 根据 240 篇查询笔迹的前 M 选候选名单中是否出现查询笔迹的实际书写者, 可以计算出笔迹鉴别的前 M 选正确率. 而 $M = 1$ 时, 就得到首选正确率. 同样, 对 93 人的英文笔迹库、266 人的藏文笔迹库和 120 人的维吾尔文笔迹库, 也得出我们方法的鉴别正确率. 由于汉字字符笔画比较密集, 笔迹笔画相对比较短促, 我们在提取微结构特征时取 $n = 7$, 即选取考察窗口大小为 15×15; 而对英文、藏文和维吾尔文笔迹, 则取 $n = 10$, 窗口大小为 21×21. 表 1 ∼ 4 分别给出了基于微结构特征的方法在汉字、英文、藏文和维吾尔文笔迹鉴别上的首选、前 5 选和前 10 选正确率, 并比较了不同距离度量方法的性能. 最好的汉字笔迹鉴别首选正确率达到 94.6 %, 而英文、藏文和维吾尔文笔迹鉴别的首选正确率分别达到 98.9 %、92.5 % 和 91.7 %. 需要说明的是, 汉字笔迹库中同一人的参考笔迹和查询笔迹原本属于同一篇文本, 文本内容本身有一定相

表 1 汉字笔迹鉴别正确率 (240 人)

Table 1 Identification accuracy rates of Chinese handwritings (240 writers)

距离度量	首选正确率 (%)	前 5 选正确率 (%)	前 10 选正确率 (%)
欧氏距离	78.3	93.8	95.4
卡方距离	87.5	95.4	96.7
加权欧氏距离	92.5	97.1	97.9
加权卡方距离	94.6	97.5	98.8

表 2 英文笔迹鉴别正确率 (93 人)

Table 2 Identification accuracy rates of English handwritings (93 writers)

距离度量	首选正确率 (%)	前 5 选正确率 (%)	前 10 选正确率 (%)
欧氏距离	93.6	96.8	98.9
卡方距离	95.7	98.9	98.9
加权欧氏距离	98.9	100	100
加权卡方距离	97.9	100	100

表 3　藏文笔迹鉴别正确率 (266 人)

Table 3　Identification accuracy rates of Tibetan handwritings (266 writers)

距离度量	首选正确率 (%)	前 5 选正确率 (%)	前 10 选正确率 (%)
欧氏距离	79.0	88.0	90.6
卡方距离	83.5	89.5	91.0
加权欧氏距离	92.1	95.5	96.6
加权卡方距离	92.5	95.9	97.4

表 4　维吾尔文笔迹鉴别正确率 (120 人)

Table 4　Identification accuracy rates of Uighur handwritings (120 writers)

距离度量	首选正确率 (%)	前 5 选正确率 (%)	前 10 选正确率 (%)
欧氏距离	83.3	90.8	91.7
卡方距离	85.8	92.5	92.5
加权欧氏距离	91.7	96.7	98.3
加权卡方距离	91.7	96.7	97.5

表 5　各种特征在 IAM 英文笔迹库上的鉴别正确率 (留一法、卡方距离)

Table 5　Identification accuracy rates of different features on IAM English handwriting database (Leave-one-out, χ^2 metric)

特征	人数	首选 (%)	前 5 选 (%)	前 10 选 (%)
Contour-Hinge[11]	650	81	N/A	92
Grapheme Emission[11]	650	80	N/A	94
微结构特征	657	85.3	93.0	94.2

关度, 而且书写者在同一篇文本中书写差异也会相对较小. 所以, 虽然汉字笔迹样本的字符数相对较少, 但是在汉字笔迹库上的鉴别正确率比藏文和维吾尔文笔迹都要高. 而英文笔迹库中书写人数相对较少, 且不同书写人的文本内容有一定的重复情况, 因此鉴别正确率在数值上表现得最好.

为了与文献 [11] 中的特征进行比较, 我们在 IAM 全部 657 人的手写文本上进行了笔迹鉴别实验, 将其中部分书写人的单篇文本进行如同 HIT-MW 库一样的分割处理, 并采用与文献 [11] 相同的卡方距离度量和 "留一法 (Leave-one-out)" 实验策略. 表 5 给出了使用网格窗口微结构特征的笔迹鉴别性能与文献 [11] 中结果的比较. 这里仅列出了文献 [11] 中两种表现最好的特征, 即 Contour-Hinge 特征和 Grapheme emission 特征. 虽然文献 [11] 只使用了 650 个书写人, 而我们使用了全部 657 人, 但是可以看出, 我们提出的微结构特征在鉴别性能上要优于另两种特征.

4　结论

从实验结果可以看出, 基于网格窗口微结构特征的笔迹鉴别方法适用于多种文种的笔迹鉴别, 并且在较大规模的笔迹样本库上都取得了很好的鉴别性能. 这一结果也证明了我们的分析, 微结构特征可以反映出不同书写者笔迹风格的差异, 也能够适应不同文种笔迹并不相同的局部细微结构, 并对其加以描述. 从实验结果还可以看出, 使用加权版本距离度量的性能均优于使用原始距离度量的性能, 而且两种加权版本距离度量的性能基本相当. 这说明, 用各维分量标准差对元素值进行加权, 能够有效提高我们笔迹鉴别方法的性能. 因为不像其他基于纹理分析的笔迹鉴别方法需要构造文本纹理块和进行多通道滤波器滤波, 我们提出的基于微结构特征的笔迹鉴别方法直接对笔迹图像进行处理, 并且直接考察图像的实际像素值, 所以基于微结构特征的笔迹鉴别方法在实际应用上更有优势. 而且网格窗口微结构特征的鉴别性能优于 Contour-Hinge 特征和 Grapheme emission 特征, 也不需要 Grapheme emission 特征的聚类学习步骤. 由于本文所提出的方法是一种文本无关的多文种笔迹鉴别方法, 不需要利用文本内容信息, 避免了对手写文本图像进行切分和文本标记, 相对于连写拼音文字笔迹而言, 更是一种实用的笔迹鉴别方法. 微结构特征的改进以及与其他特征的融合, 是下一步可能的工作.

References

1 Plamondon R, Lorette G. Automatic signature verification and writer identification — the state of the art. *Pattern Recognition*, 1989, **22**(2): 107−131

2 Li X, Ding X Q, Wang X L. Semi-text-independent writer verification of Chinese handwriting. In: Proceedings of the 11th International Conference on Frontiers of Handwriting Recognition. Montreal, Canada: IEEE, 2008. 100−105

3 Yoshimura I, Yoshimura M. Writer identification based on the ARC pattern transformation. In: Proceedings of the 9th International Conference on Pattern Recognition. Rome, Italy: IEEE, 1988. 35−37

4 Wang X L, Ding X Q, Liu H L. Writer identification using directional element features and linear transform. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, UK: IEEE, 2003. 942−945

5 Said H E S, Tan T N, Baker K D. Personal identification based on handwriting. *Pattern Recognition*, 2000, **33**(1): 149−160

6 Zhu Y, Tan T N, Wang Y H. Biometric personal identification based on handwriting. In: Proceedings of the 15th International Conference on Pattern Recognition. Barcelona, Spain: IEEE, 2000. 797−800

7 He Z Y, Tang Y Y. Chinese handwriting-based writer identification by texture analysis. In: Proceedings of the 3rd International Conference on Machine Learning and Cybernetics. Shanghai, China: IEEE, 2004. 3488−3491

8 He Z Y, Fang B, Du J W, Tang Y Y, You X G. A novel method for offline handwriting-based writer identification. In: Proceedings of the 8th International Conference on Document Analysis and Recognition. Seoul, Korea: IEEE, 2005. 242−246

9 He Z Y, You X G, Tang Y Y. Writer identification of Chinese handwriting documents using hidden Markov tree model. *Pattern Recognition*, 2008, **41**(4): 1295−1307

10 Bulacu M, Schomaker L, Vuurpijl L. Writer identification using edge-based directional features. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, UK: IEEE, 2003. 937−941

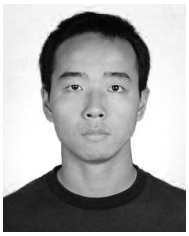
11 Bulacu M, Schomaker L. Text-independent writer identification and verification using textural and allographic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(4): 701−717

12 Bulacu M, Schomaker L, Brink A. Text-independent writer identification and verification on offline Arabic handwriting. In: Proceedings of the 9th International Conference on Document Analysis and Recognition. Curitiba, Brazil: IEEE, 2007. 769−773

13 Pratt W K. *Digital Image Processing*. New York: Wiley, 1991

14 Su T H, Zhang T W, Guan D J. Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text. *International Journal of Document Analysis and Recognition*, 2007, **10**(1): 27−38

15 Marti U V, Bunke H. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal of Document Analysis and Recognition*, 2002, **5**(1): 39−46



李 昕 清华大学电子工程系博士研究生. 2003 年获清华大学电子工程系学士学位. 主要研究方向为模式识别与生物特征识别. 本文通信作者. E-mail: lixin@ocrserv.ee.tsinghua.edu.cn

(**LI Xin** Ph. D. candidate in the Department of Electronics Engineering, Tsinghua University. He received his bachelor degree from Tsinghua University in 2003. His research interest covers pattern recognition and biometrics identification. 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, 生物特征识别和视频监控.
E-mail: dxq@ocrserv.ee.tsinghua.edu.cn

(**DING Xiao-Qing** Professor in the Department of Electronics Engineering, Tsinghua University. Her research interest covers image processing, pattern recognition, biometric identification and authentication, and video surveillance.)

彭良瑞 清华大学电子工程系讲师. 主要研究方向为图象处理、模式识别、文档分析与识别和数字信号处理.
E-mail: plr@ocrserv.ee.tsinghua.edu.cn

(**PENG Liang-Rui** Lecturer in the Department of Electronics Engineering, Tsinghua University. Her research interest covers image processing, pattern recognition, document analysis and recognition, and digital sinal processing.)