基于图像结构与汉字形态特征的印鉴自动检索方法

刘 宏[1] 宋 勇[2]
(北京大学视觉与听觉信息处理国家重点实验室 北京 100871)[1]
(北京大学金融信息化研究生课程进修班 北京 1000871)[2] (北京大学信息科学中心 北京 100871)

摘 要 本文从印鉴的图像结构与汉字形态特征分析入手,提出了一种支持印鉴图像自动检索的方法。首先,利用印鉴边框的特点快速准确求取印章的中心位置,利用章面的对称性确定主文字区域,根据文字分布的均匀性实现汉字的分割。其次,从纵横两个方向测量单个汉字的外围轮廓,形成描述汉字形态特征的位置码和长度码,进而构成支持印鉴图像检索的汉字串形态特征码。最后,通过实验分析了上述方法的可行性。

关键词 印章,图像结构,图像检索,汉字形态

1. 引言

印章作为法人和个人身份的鉴别手段,在中国已经有几千年的历史。现代科技高度发达的今天,印章仍然在中国、日本和韩国等地区被广泛使用。传统的人工鉴别已经难以满足印章管理和鉴别的要求,通过计算机自动处理印鉴图像、鉴别印鉴真伪成为具有重要的理论意义和应用价值的研究课题。

印章鉴别是图像处理和模式识别领域一个比较典型的问题,国内外学者已经开展了比较深入的研究,特别是日本学者在该领域的研究工作开展较早[1~3]。Ueda 和 Nakamura[3]根据人工折角比对的方法,开发了基于若干局部和全局特征的统计决策方法。Fan 和 Tsai[4]根据印鉴笔划的拓扑结构具有的相对稳定性,提出基于细化后骨架结构的识别方法。Lee 和 Kim[5]引入笔划特征图 ASG,把关系图和几何位置结合起来进行判决,该方法突破了对印鉴形状和拓扑结构完整性的限制。胡庆和杨静宇等[6~8],从知识工程的角度出发,引入差图像的概念,并结合统计和结构模式识别的方法,提出一种基于多特征的鉴别方法。高文等[9]在考虑印鉴中笔锋、笔划粗细及笔划相对位置等特征的基础上,提出了一种基于边缘匹配的鉴别方法,利用边缘点的距离作为识别特征。这些研究工作主要研究印鉴图像的自动真伪判别问题和自动旋转匹配问题,标准印鉴是通过人工输入相应的 ID 号码提取的,适合于一对一的自动鉴别情况,但无法处理一对多的自动识别。

利用印鉴中的图像内容建立索引,摆脱对人工输入印章 ID 号的依赖,是实现全自动印鉴识别的关键。但采用汉字识别技术处理印鉴检索问题,存在着一定的困难:首先在公用印章图像中,文字呈弧线排列在章体内部,其提取需要根据章面的布局和结构信息,汉字识别中一般的汉字检测和分割方法是难以奏效的;其次,印鉴图像中的文字与印刷体和手写体都有明显的差别,文字加盖的效果受到油墨多少、加盖时用力大小、印油渗透情况、图文背景干扰等多种因素的影响,加盖模糊和部分印痕残缺的现象是很常见的,对汉字笔划和结构的分析造成很大的干扰。但通过观察发现,在加盖效果不理想的印鉴图像中,汉字的形态特征具有良好的稳定性,通过多个汉字形态特征的排列,汉字串的形态特征将具有更强的区分能力,可望成为有效的印鉴图像检索手段。

2. 图像结构特征与文字区域分割

印章的使用与中国汉字特殊的结构有密切关系。章面上刻有多个汉字,按弧线或直线很规则地排列,具有内容直观、内涵丰富、颜色鲜明、庄严朴实的特点。以本文处理的公章印鉴为例,其印鉴图像的结构特征包括以下五个方面,如图 1 所示。

1)印章具有圆形的边框,而边框宽度均匀,形态鲜明;

2)印章在整体布局上呈左右对称形式,存在一条结构上对称的主轴;

3)章面上主文字区域以印章中心为圆心,沿着圆弧呈扇形均匀分布;

4)章面上辅文字区域(可能缺省)呈直线均匀分布,与对称主轴垂直;

5)五角星位于印章中心,其中心和一个顶点位于印章的对称主轴上。

利用上述结构特征,在二值化的印鉴图像中提取的主文字区域的汉字,具体步骤包括:

1)印章中心位置的确定:去除噪声干扰后计算印鉴图像的重心 G,通过 G 点每隔 45 度划一条直线,根据印章边框与重心 G 的距离、边框宽度等信息确定这些条直线与边框的八个交点,记为 C_i,如图 2 所示,这里 i=1,2,…,8。对于残缺印鉴或有强

背景干扰的印鉴，求得的这 8 个交点未必都在印章的边框上。但从原理上讲，只要有三个交点确实在边框上，就可以比较准确地求出印章的中心位置，即五角星的中心位置，记做 M。由于五角星部分加盖效果往往很不均匀，因此不能采用简单求取五角星区域中心的方法计算印章中心。

图 1　本文实验用印章均在法定印章制作单位按标准刻制

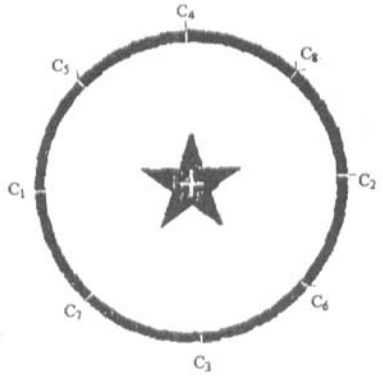

图 2　边框上的 8 个交点与印章中心

图 3　二值化后的图像及文字区域采样线

2)主文字区域的定位与提取：在确定了印章中心位置 M 后，根据印鉴图像中主文字区域与印章中心的距离范围和围绕印章中心按圆弧排列的特点，以 M 点为中心，设计 10 条穿过文字区域的圆作为采样线，如图 3 所示。将采样线上的点记入数组 A_{ij}，这里，i=1,2,…,10;j=1,2,…,360。按 i 值接近的原则将 A_{ij}的元素分为三组并累加起来，分析采样线上像素数量的变化，结合印章对称性确定主文字区域的起始和结束位置，如图 4 所示。

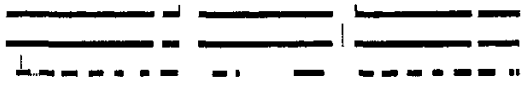

图 4　根据采样线上汉字像素数量的变化确定主文字区域

3)汉字的分割：确定了主文字区域的起始和结束角度、与 M 点的距离范围后，将主文字区域采集下来，记为数组 Z_{ij}，累加数组的每一列上所有行的像素点，记为 W_i;计算 W_i 与 W_{i+k}的相关性，在 k 的不同取值位置形成多个峰值，通过对这些峰值的检测，求取文字间隔距离 H，如图 5 所示，图中的直角短线标出了峰值的位置。根据文字间隔距离 H 并结合局部的动态调整，完成对主汉字区域的精确分割，如图 6 所示。图 6 中还标出了每个文字的 v_1 和 v_2 两个形态特征在文字图像中的位置。

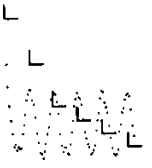

图 5　主文字区字符间距的分析

图 6　主文字区字符分割结果

3. 支持印鉴检索的汉字形态特征

汉字是中华民族五千年文化的结晶，具有丰富的内涵。除字意、发音这两大属性外，汉字还具有很强的形态属性。这里定义单个汉字的“横肩”、“横

胯”、“横胫”、“横腰”、“纵肩”、“纵胯”、“纵胫”、“纵腰”等8种主要形态特征，具体定义如下：

“横肩”：字体横向最宽的部位，记作 v_1；

“横胯”：字体横向次宽的部位，记作 v_2；

“横腰”：字体横向次窄的部位，记作 v_3；

“横胫”：字体横向最窄的部位，记作 v_4；

“纵肩”：字体纵向最宽的部位，记作 h_1；

“纵胯”：字体纵向次宽的部位，记作 h_2；

“纵腰”：字体纵向次窄的部位，记作 h_3；

“纵胫”：字体纵向最窄的部位，记作 h_4；

这里，以“新”和“年”两个字的字体为例，说明上述特征及其位置，如图7所示。

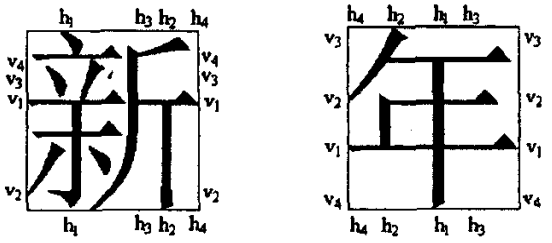

图7 本文定义的汉字形态特征示意图

8种形态特征的位置 $v_1, v_2, v_3, v_4, h_1, h_2, h_3, h_4$ 和对应长度数据 $lv_1, lv_2, lv_3, lv_4, lh_1, lh_2, lh_3, lh_4$ 可以构成两个向量，分别称为汉字形态特征位置码和长度码，即：

位置码：$\mathrm{Pcode}(v_1, v_2, v_3, v_4, h_1, h_2, h_3, h_4)$

长度码：$\mathrm{Lcode}(lv_1, lv_2, lv_3, lv_4, lh_1, lh_2, lh_3, lh_4)$

常用汉字有6000多个，根据位置码中8种特征出现次序的不同来排列，可以将汉字分为 $24\times24=576$ 类，平均每个类包含十几个汉字。当然，这只是平均意义上的数量。印章文字中的汉字组合受到语法和语意的双重约束，因此，在汉字形态特征上非常接近的比较长的汉字串是比较少见的。但是，这种判断还要基于对大量的印章文字形态的统计研究，这也是目前我们正在开展的一项研究工作。尽管这种编码方法不能唯一地描述一个汉字，但存在着很多有价值的特点：

1)特征直观形象，既适合计算机处理，又便于肉眼观察。

2)特征提取对汉字结构依赖性小，可以容忍印鉴图像加盖的模糊性。

3)易于提取：只需要简单的加法和比较运算。

4)特征描述简单，信息量大：既可以利用距离信息，又可以利用排列信息，如纵横特征的总体排列、纵向(横向)位置码的内部排列等。这样，可以利用纵横关系、特征排列和特征取值三个层次的信息，支持多级检索。

5)可以处理倒置或横置的成像文档：由于汉字形态上方方正正的特点，在缺乏明显标记和正常排版信息的情况下，确定文档阅读的正方向是比较困难的，而该方法提取特征的过程与文档是否倒置或横置无关，而斜置的情况是很容易发现和矫正的。

6)适合多种字体：不同的字体在笔划粗细和走向上有明显变化，但在汉字的形态特征上却具有比较强的稳定性，适合在未知字体的情况下实现文字检索。

7)组合特征的检索性能与汉字串长度成正比，而印章中文字数量一般较多(多数印章中主文字区域的字数在十几个以上)。多个连续汉字的形态特征排列在一起，就组成了汉字串的形态码，具有比单个汉字更强的特征区分能力。

8)形态特征不依赖于对汉字的识别：识别印鉴上的每个汉字来达到印鉴图像检索的目的，要付出比分析形态特征更高的代价：首先要建立很大的汉字特征库，因为印章中可能出现的汉字组成的字符集是非常大的；另外，由于印鉴加盖过程中各种复杂因素的影响，印鉴图像中的文字往往是不规范的，与印刷体和手写体存在很大的不同，采用一般的汉字的识别方法，难以取得满意的识别结果。

9)形态特征可应用于更广泛的领域，例如：在成像文档的关键词检索中，可以通过相似性筛选，大幅度降低可疑词的范围，减少识别处理的工作量，大大提高检索速度。

4. 形态特征提取与自动检索策略

根据第3节给出的汉字形态特征的定义，计算汉字形态特征的过程和基本方法如下：

1)形态特征提取：采用隔行(列)扫描加快提取速度，通过各行(列)像素的累加值的比较，确定形态特征的位置码和长度码。根据汉字笔划的连通性和宽度信息去除孤立噪声对特征提取的干扰。利用两个相邻特征位置的距离限制，避免特征的重复提取。

2)置信度分析：对形态特征位置码进行置信度分析，作为辅助判别信息，可以有效处理“肩”和“胯”之间、“腰”和“颈”之间的错判问题。

3)汉字串形态特征码的形成：位置码、长度码和置信度信息构成了单个汉字形态特征的描述信息。将多个连续汉字的形态特征信息排列在一起，就构成了支持印鉴检索的汉字串形态特征码。

印鉴图像自动检索既涉及印鉴结构特征，又涉及汉字形态特征。本文提出的检索策略是：以结构特征支持高层检索，以形态特征支持低层检索，按层次组织汉字串形态码，支持快速检索。具体检索步骤为：

1)结构检索：首先根据主文字区域的扇面角度和文字数量分类。

2)定性检索：利用汉字串形态特征码中的位置

码及其置信度信息判断两个汉字串的相似性。

3)定量检索:利用汉字串形态特征码中的长度码判断两个汉字串的相似性,这一定量检索允许与标准的特征取值存在一定的误差,甚至个别特征取值匹配的失败。

通过以上三级处理找到匹配的汉字串特征码,则检索成功。

5. 实验与结论

为检验本文提出的印鉴图像自动检索方法的可行性,建立了包括百枚印鉴图像的数据库,通过数据模拟了上万枚印鉴图像的汉字串形态特征库。在万枚规模的特征库中,可以实时地检索印鉴图像;在百枚规模的印鉴图像库中可以检索出待测印鉴,正确率达到 90%以上;未正确检出的,均为拒识的情况,未发生错误检索的情况。初步的实验表明:该方法不仅能处理加盖清晰完整的印鉴,还可以处理加盖比较模糊、残缺的印鉴,也可以处理具有复杂图文背景干扰的印鉴,如图 8 至图 10 所示,但检索的正确率比清晰完整印鉴要低 15%左右。

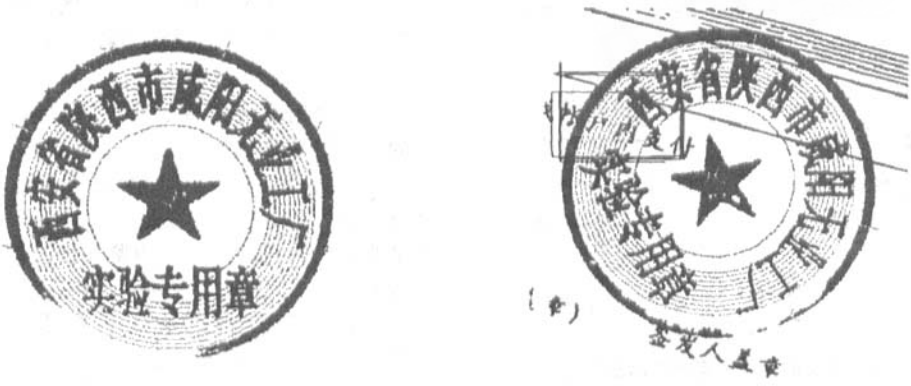

图 8 实现自动检索的印鉴图像:残缺印鉴和有背景干扰的印鉴

图 9 采样线上汉字像素数量的变化(对应于图 8 中的两个印鉴)

图 10 主文字区字符间距分析和字符分割结果(对应于图 8 中的两个印鉴)

结论 本文提出了一种基于图像结构和汉字形态的印鉴自动检索方法。在分析印章结构的基础上,给出了主文字区域的定位、提取和汉字分割方法;定义了 8 种主要的汉字形态特征,给出了形态特征提取方法和实现快速检索的策略,并通过实验初步检验了这些方法的可行性。

尽管目前印鉴图像数据库规模还比较小,但图像结构特征的分析和处理方法与实验数据规模没有直接关系;而形态特征的快速提取和分层次检索策略保证了这种方法可以支持大型检索系统的实时性要求。印鉴图像检索性能的提高主要取决于汉字串形态特征码对汉字串的区分能力,通过增加新的容易提取而又个性丰富的形态特征将是提高汉字串区分能力,提高印鉴图像检索性能的有效途径。

参 考 文 献

1 Mieno H. An experiment of identification of seal-impression by pattern matching. Information Processing, 1975, 16(3)

2 Kaneko T. Positioning of seal-impressions using marginal densities about the centroid. Trans. of the IECE, 1984, J67-D(1)

3 Ueda K, Nakamura Y. Automatic verification of seal-impression patterns. In: Proc. 7th Int. Conf. PR, 1984. 1019~1021

4 Fan T J, Tsai W H. Automatic Chinese seal identification. CVGIP, 1984, 25: 311~330

5 Lee S, Kim J H. Unconstrained seal imprint verification using attributed stroke graph matching, Pattern Recognition, 1989, 22: 653~664

6 胡庆,杨静宇,等. 基于知识的印鉴鉴别方法. 自动化学报,1991, 17(6): 696~704

7 张黔,胡庆,杨静宇,等. 统计和结构模式识别相结合的多特征印鉴真伪鉴别方法. 计算机学报,1995,18(3): 190~198

8 Hu Q, et al. An automatic seal imprint verification approach. Pattern Recognition, 1995, 28(8): 1251~1266

9 高文,等. 基于边缘匹配的印鉴自动鉴别方法. 模式识别与人工智能,1994,7(4): 338~342