综合文字和非文字区域特征的文档图像检索

张 田
ZHANG Tian

山东大学 信息科学与工程学院,济南 250100
School of Information Science and Engineering,Shandong University,Jinan 250100,China

ZHANG Tian.Document image retrieval method using combination of text and non-text features.Computer Engineering and Applications,2010,46(12):5-8.

Abstract: An improved self-adaptive method for text area extraction is proposed.With it,the document image is segmented into text area and non-text area firstly.And then,for text area,local features and global features are extracted.The local features include gaps between connected characters,height and width of connected characters,and the global features contain writing style and paragraph features.For non-text area,the key block feature is extracted.After that,the retrieval method combines all the features to improve the accuracy.Meantime,multi-dimensional retrieval structure is introduced to improve the speed.The experiments performed on a large-scale document image database (including 12,024 images) reveal that the method is more efficient than existing ones.
Key words: document image retrieval;text area extraction;paragraph feature;multi-dimensional retrieval structure

摘 要:提出一种改进的自适应文字区域提取算法,将文档图像分割成文字区域和非文字区域。对文字区域提取连通字符间空白、连通字符高度和宽度等局部特征,以及书写样式、段落特征等全局特征;对非文字区域,提取关键块特征。然后利用检索算法将文字区域特征和非文字区域特征结合起来,提高检索的准确性。同时,在检索算法中引入多维数据检索结构,有效地提高检索速度。通过对大规模文档数据库(包含 12 024 个文档)的检索,表明该算法具有较高的效率,优于现有的一般文档图像检索算法。
关键词:文档图像检索;文字区域提取;段落特征;多维数据检索结构
DOI:10.3778/j.issn.1002-8331.2010.12.002 **文章编号**:1002-8331(2010)12-0005-04 **文献标识码**:A **中图分类号**:TP391

1 引言

文档图像检索是图像检索的重要方面，在数字图书馆、档案管理、办公自动化等方面有广泛的应用。文档图像检索一般是指从文档图像数据库中找出与输入文档图像相匹配的图像。常见的文档图像检索算法可归为基于文档字符内容检索和基于图像特征检索两类[1]。基于图像特征检索具有时间复杂度低、适应性强的特点,它依靠某些文档图像本身的特征实现文档图像的检索。

典型的基于特征的文档图像检索算法包括:P.Herrmann 等给出的基于文档的版面特征的检索[2]。H.Peng 等给出的基于文档图像中段落块系列大小和位置信息的匹配算法[3]。D.Doermann 等提出一种先提取文档图像的代表行作为文档图像签名,然后获取该行的字符形状编码,最后利用这些字符形状编码去检索图像的算法[4]。C.L.Tan 等提出的基于字符对象的水平横断密度和垂直横断密度的方法[5]。C.Wang 等提出的基于字符外接最小矩形范围内前景像素比例的检索方法[6]。胡芝兰等提出的通过提取有效文本区域的长宽比和分层密度特征,然后通过特征比对进行检索的方法[7]。H.Liu 等提出的通过文档图像的密度分布特征和关键块特征进行检索的方法[8]。G.F.Meng 等给出的基于多个文档图像特征综合使用的文档图像检索方法[9]。

这些方法中有的是基于文档图像的字符特征等局部特征的(如文献[4-6]),容易受畸变、噪声、扫描品质、分辨率等因素的影响;有的是基于文档图形的版面特征等全局特征的(如文献[2-3]),与局部特征相比,这些全局特征对图像分辨能力又较弱[8]。文献[7]虽然用到了全局特征和局部特征,但仅仅是分离使用,检索算法在第一步中使用长宽比进行初步检索,在第二步中使用分层密度特征进行进一步的检索,而未能将两者充分结合起来。文献[8]明确提出了使用全局特征和局部特征的文档图像检索，检索算法在第一步中使用密度分布特征行检索给出 10 个最佳匹配，在第二步中使用关键块特征对第一步的结果进行进一步筛选给出 5 个最佳匹配。但还是分离使用,而未能给出有效算法将两者充分结合起来。文献[9]给出的方法较好地实现了多个文档图像特征的综合使用,但算法从文档图像特征库中检索与输入特征最相似的多个特征,进而给出候选图像集合的过程中使用的是顺序搜索,特征库规模较大时,搜索速度较慢。

2 研究工作

该文的工作针对的是一般文档图像,既可以是印刷体也可

基金项目:国家自然科学基金重点项目(the Grand National Natural Science Foundation of China under Grant No.60832008)。
作者简介:张田(1984-),男,博士研究生,研究方向:信号处理,无线通信。
收稿日期:2010-01-27 **修回日期**:2010-03-12

以是手写体,或者两者混合,包括:

(1)提出一种改进的文字区域提取算法,将文档图像的文字区域和非文字区域分离。对文字区域,提取字符间空白、字符高度和宽度等局部特征以及书写样式、段落特征等全局特征,其中提出基于数学形态学的算法用于段落特征的提取;对非文字区域提取关键块特征。

(2)通过联合组合给出检索方法,提高了检索的精度。引入多维检索结构,有效地提高检索速度。

(3)组织、建立了一个大规模的文档图像数据库(包含12 024 个文档),并通过在该数据库上的实验验证了算法的有效性。

3　文档图像特征提取

3.1　文字区域、非文字区域分割

嵌入在文档图像中的文字信息是图像语义的一种重要表达方式,通常反映了图像的主要内容。文中将文档图像的文字区域和非文字区域分割,分别提取特征。

在文献[10]的基础上引入 LMS 算法[11]提出一种改进的自适应算法。自适应地给出最佳分割 ξ。

输入:图像对序列即样本$<S_i,I_i>$,$i=1,2,\cdots,M$,S_i 为原图像,I_i 为对应分割结果。要处理的图像记为 Q。

输出:映射 ξ_{opt},分割结果 $\xi_{opt}(Q)$。

步骤如下:

(1)初值 $\xi=\xi_0$

ξ_0 由 ISI 算法[12]通过训练样例学习得到[10]。

(2)迭代

$$e_i(n)=|I_i-\xi_n(S_i)|$$

$$e(n)=\sum_{i=1}^{M}e_i(n)$$

$$\xi_{n+1}(S_i)=\xi_n(S_i)+2\mu e_i(n)S_i$$

$$\xi_{n+1}(Q)=\xi_n(Q)+2\mu\frac{e(n)}{M}Q,\mu>0\text{ 为常数}$$

(a)原图像　(b)分割结果

图 1　用于训练的样例图像对

(a)待分割图像　(b)分割结果

图 2　文字与图形、表格的分割

如果 $\sum_{i=1}^{M}[\xi_{n+1}(S_i)-\xi_n(S_i)]<\theta$,停止迭代,输出 $\xi_{opt}=\xi_n$,$\xi_{opt}(Q)=\xi_n(Q)$。

3.2　文字区域的特征提取

对于文字区域,提取局部特征和全局特征。

(1)局部特征

①连通字符之间的距离,即空白 gap_i。

②连通字符的高度 h_i、宽度 w_i。

图 3　连通体之间的空白

(2)全局特征

①连通体的个数 n_1 和空穴的个数(封闭的区域数)n_2。这一特征可以获得书写样式(草书还是楷体)特征[13]。一般草书一个连通字符包含多个空穴,而楷体一个连通字符包含少于 1 个空穴。

②文字区域的段落特征。

该文提出了基于数学形态学的文档图像段落检测方法[14]。步骤如下:

(1)确定膨胀模板 $T(l,d)$。由字符平均高度 h 与行平均间距 Lg,取 $d=h+Lg$,膨胀后段落标记高度约为一个行间距。设同一行中字符(串)之间的平均间距为 Cg,取 $l=3Cg$。

(2)进行膨胀运算,得到段落标记。假设文档图像经二值化后为 $I(i,j)$,膨胀运算后为 $R(i,j)$,则 $R(i,j)$可表述如下:

$$R(i,j)=\begin{cases}0,\text{sum}(D(I(i,j),T(d,l)))<N\\1,\text{others}\end{cases}$$

N 取值为 1。$\text{sum}(D(I(i,j),T(d,l)))$表示将区域 $D(I(i,j),T(d,l))$(见图 4)内的所有像素值累加。

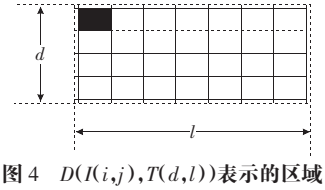

图 4　$D(I(i,j),T(d,l))$表示的区域

(黑方格表示当前像素 $I(i,j)$)

(3)记录段落标记的平均像素数目为 $Blen$,最大 $B\max$,最小 $B\min$。段落数目为 Bn,段落特征 $Block=(Bn,\text{Blen},B\max,B\min)$。

定义向量

$$\boldsymbol{word}=(gap_1,gap_2,\cdots,gap_{n_1-1},h_1,\cdots,h_2,\cdots,h_{n_1},w_1,w_2,\cdots,w_{n_1},n_1,n_2,Block)$$

为文字区域特征向量。

3.3　非文字区域的特征提取

文档图像关键块特征是文献[8]提出的一种反映文档图像页面关键几何组成信息的特征。是针对整个文档图像提取的特征。这里限定在非文文字区域。关键块一般情况下都是图片、表格等非文字区域[8],所以这种限定节约了处理时间,同时对检索影响又不大。

如图 5,图中 C_1、C_2、C_3 分别标明文档图像的第一、第二、第三个关键块前景像素的质心。定义向量 $\boldsymbol{KBF}=(kb_1,kb_2,kb_3,\vartheta_1,$

$\partial_2,\partial_3,\ell_1,\ell_2,\ell_3$)为关键块特征向量[8]。其中分量 kb_1、kb_2、kb_3 依次代表文档图像前 3 个关键块前景色带的平均宽度的相对大小;∂_1、∂_2、∂_3 依次代表着文档图像前 3 个关键块前景色带的质心的相对位置;ℓ_1、ℓ_2、ℓ_3 依次代表着文档图像前 3 个关键块内容的归类[15](表格=0、图片=1)

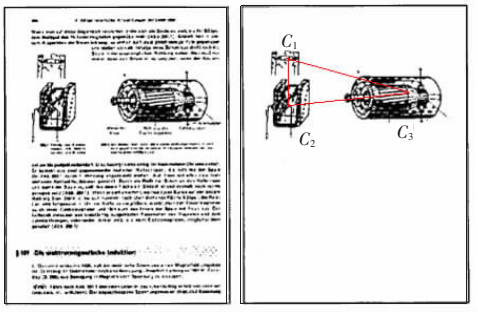

图 5　关键块特征

4　检索

检索采用联合组合(Union Combination)[9,16]的策略将提取的特征充分结合,针对大规模文档图像,引入多维检索结构 A-Tree[17]。

分别以 ***word***、***KBF*** 建立索引结构 A-Tree,记做 A_1、A_2。

4.1　构造候选图像集合

算法步骤如下:

(1)计算查询样本图像 *sample* 的 ***word***、***KBF*** 特征向量,分别记作 $F_1(sample)$、$F_2(sample)$。

(2)以 $F_1(sample)$为输入对 A_1 做 m_1 近邻查询得到与输入特征最相近的 m_1 个向量,设与之对应的文档图像集合为 Set_1。同样用 $F_2(sample)$为输入对 A_2 做 m_2 近邻查询得到与输入特征最相近的 m_2 个向量,设与之对应的文档图像集合为 Set_2。

$$Set_1=\{image_1^1,image_1^2,\cdots,image_1^{m_1}\}$$

$$Set_2=\{image_2^1,image_2^2,\cdots,image_2^{m_2}\}$$

(3)给出候选集合 *Set*:

$$Set=Set_1\cup Set_2=\{image^1,image^2,\cdots,image^t\}$$

$$\max(m_1,m_2)\leqslant t\leqslant m_1+m_2$$

4.2　从候选集合中得到查询结果

记集合 Set_1、Set_2 的权值为 w_{set}^1、w_{set}^2。$W_1=\{w_1^1,w_1^2,\cdots,w_1^{m_1}\}$、$W_2=\{w_2^1,w_2^2,\cdots,w_2^{m_2}\}$分别为集合 Set_1、Set_2 中对应图像的权值。$W=\{w^1,w^2,\cdots,w^t\}$为候选集合 *Set* 中对应图像的权值。取 $w_1^i=\delta_1(m_1-i+1),i=1,2,\cdots,m_1$,$w_2^i=\delta_2(m_2-i+1),i=1,2,\cdots,m_2$。$\delta_1,\delta_2>0$ 为集合的初始系数,用来表征集合的初始重要程度。候选集合中图像的权值是该图像在两个集合中对应权值的累加,即:

$$w^i=\sum_{image^i\in Set_j\&image^i=image_j^k,j=1,2}w_j^k,i=1,2,\cdots,t \qquad (1)$$

从候选集合中得到查询结果的步骤如下:

(1)根据权值对候选集合 *Set* 中的图像做降序排列。得有序的候选集合 $S=\{im^1,im^2,\cdots,im^t\}$,以及对应权值集合 $W_{se}=\{w_{se}^1,w_{se}^2,\cdots,w_{se}^t\}$。

(2)选择有序候选集合 *S* 中的前 *m* 个图像。*m* 为正整数。

(3)计算集合 Set_1、Set_2 的权值 w_{set}^1、w_{set}^2。对于有序候选集合 *S* 中的前 *m* 个图像如果其是某个集合中的图像则将其在该集合中的权值加到该集合的权值中。即

$$w_{set}^i=\sum_{im^k\in Set_i\&im^k=image_i^j}w_i^j,i=1,2$$

(4)调整 Set_1、Set_2 中对应图像的权值 W_1、W_2。

$$W_i=w_{set}^i\times W_i,i=1,2$$

(5)按照式(1)重新计算候选集合中图像的权值 w^i,依照最新的权值重新降序排列候选集合 *Set* 中的图像,取前 *m* 个作为查询结果。

5　实验及分析

5.1　实验数据

为了测试该方法的效果,首先构造一个文档图像数据库,图像采集使用山东山大欧玛软件有限公司的 MOS 770 高速扫描设备,共包含 12 024 张不同分辨率(100 dpi(dot/inch),200 dpi,400 dpi)的 256 级灰度图像,有主要是文字的,还包括文字、图片、表格混合的。如图 6。

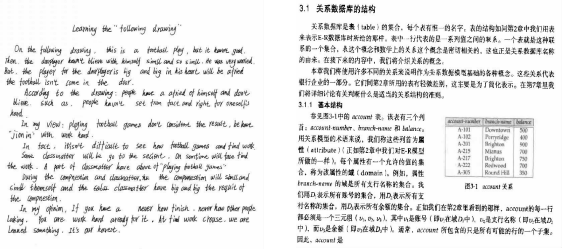

图 6　测试用例

ISI 算法的学习样例,从数据库中随机选取,然后人工标注,有 300 对。

测试样本共 200 幅,从数据库的文档图像中随机抽取得到。

5.2　性能评价指标

采用规格化的 Recall-Precision 准则[18]作为系统性能评价标准。该准则是在传统的 Recall-Precision 准则中增加了对返回结果排序的考量。

$$Recall=1-\frac{\sum_{i=1}^{n}R_i-\sum_{i}^{n}i}{n(N-n)}$$

$$Precision=1-\frac{\sum_{i=1}^{n}(\lg R_i)-\sum_{i}^{n}(\lg i)}{\lg(\frac{N!}{(N-n)!\,n!})}$$

其中 *Recall*、*Precision* 分别为查全率和查准率,R_i 为第 i 个相关图像在检索结构中的排序,n 为相关图像的个数,N 为图像库的规模(图像总数)。

5.3　实验流程

实验流程见图 7。

5.4　实验结果及分析

算法中的参数 $m_1=m_2=10,\delta_1=\delta_2=1,\mu=0.1,\theta=0.01$。

一次检索输出图像数 *m* 为 1 时的实验结果如表 1:

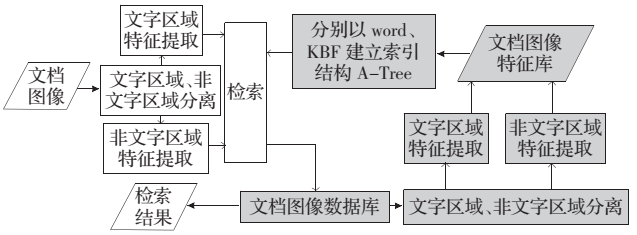

图 7 实验流程图

（灰色部分线下完成）

表 1 一次检索输出图像数 m 为 1 时的实验结果

	word 特征	KBF 特征	word、KBF 特征结合
Precision/(%)	79.8	63.0	92.0
Recall/(%)	82.0	75.1	94.3

一次检索输出图像数 m 为 6 时的实验结果如表 2：

表 2 一次检索输出图像数m为 6 时的实验结果

	word 特征	KBF 特征	word、KBF 特征结合
Precision/(%)	86.0	70.4	94.1
Recall/(%)	94.2	78.0	96.9

作为对比，给出同样条件下文献[8]的实验结果见表 3，文献[9]的实验结果见表 4。

表 3 文献[8]的实验结果

	Precision/(%)	*Recall*/(%)
m=1	89.4	93.0
m=6	92.8	96.0

表 4 文献[9]的实验结果

	Precision/(%)	*Recall*/(%)
m=1	90.2	93.2
m=6	92.7	95.8

该文算法平均检索时间如表 5。

表 5 该文算法平均检索时间

阶段	文字、非文字分离	word 提取	KBF 提取	检索	合计
时间/ms	434	496	358	68	1 356

建立索引的时间并未记入，原因是这部分在线下完成，对检索效率没有影响。

相同条件下文献[8]的平均时间为 1 320 ms，文献[9]的时间为 1 297 ms。

表 1 和表 2 表明 word 特征和 KBF 特征的结合能明显提高检索精度。

表 1、表 2 和表 3、表 4 的对比表明该文算法在检索精度上优于文献[8]和文献[9]。

在检索平均时间上，该文算法与文献[8]和文献[9]分别相差 2.7%和 4.5%。

6 结论

提出局部和全局结合的检索，提出一种改进的文字区域提取算法，将文档图像的文字区域和非文字区域分离，对它们分别提取特征。然后用联合组合的策略将提取的特征综合运用，提高了检索的准确性。通过引入多维检索结构，提高了检索速度。

参考文献：

[1] 冯所前.大规模复杂文档图像快速检索系统的研究与实现[D].北京大学，2005.

[2] Herrmann P，Schlageter G.Retrieval of document images using layout knowledge[C]//Proc 2nd ICDAR，1993：537-540.

[3] Peng H，Long F，Chi Z，et al.Document image template matching based on component block list[J].Pattern Recognition Letters，2001，22(9)：1033-1042.

[4] Doermann D，Li H，Kia O.The detection of duplicates in document image databases[J].Image and Vision Computing，1998，16(12)：907-920.

[5] Tan C L，Huang W，Yu Z，et al.Imaged document text retrieval without OCR[J].IEEE Trans PAMI，2002，24(6)：838-844.

[6] Wang C，Chen T，Chan Y，et al.Imaged document image retrieval system based on proportion of black pixel area in a character image[C]//6th ICACT，2004：25-29.

[7] 胡芝兰，林行刚，严洪.基于分层密度特征的文档图像检索[J].清华大学学报：自然科学版，2006，46(7)：1231-1234.

[8] Liu H，Feng S Q，Zha H B，et al.Document image retrieval based on density distribution feature and key block feature[C]//Proceedings of the Eighth International Conference on Document Analysis and Recognition，2005，2：1040-1044.

[9] Meng G F，Zheng N N，Song Y H，et al.Document images retrieval based on multiple features combination[C]//Ninth International Conference on Document Analysis and Recognition，ICDAR 2007，2007，1：143-147.

[10] Hirata N S T，Barbera J，Terada R.Text segmentation by automatically designed morphological operators[C]//XIII Brizilian Symposium on Computer Graphics and Image Processing，SIBGRAPI'00，Sibgrapi，2000：284.

[11] 姚天任，孙洪.现代数字信号处理[M].武汉：华中科技大学出版社，2006：56-57.

[12] Barrera J，Dougherty E R，Tomita N S.Automatic programming of binary morphological machines by design of statistically optimal operators in the context of computational learning theory[J].J Electronic Imaging，1997，6(1)：54-67.

[13] Huang C.Content-based handwritten document indexing and retrieval[D].State University of New York，2008.

[14] Zhang T.A feature-based document image retrieval method[C]//Chinese Conference on Pattern Recognition，CCPR'08，Beijing，China，2008：360-364.

[15] Lee S W，Ryu D S.Parameter-free geometric document layout analysis[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，2001，23(11)：1240-1256.

[16] Ho T K，Hull J J，Srihari S N.Decision combination in multiple classifier systems[J].IEEE Trans PAMI，1994，16(1)：66-75.

[17] Sakurai Y，Yoshikawa M，Uemura S，et al.The A-tree：An index structure for high-dimensional spaces using relative approximation[C]//Proc of the 26th International Conference on Very Large Data Bases(VLDB)，2000：516-526.

[18] 林传力，赵宇明.基于 Sift 特征的商标检索算法[J].计算机工程，2008，34(23)：275-277.