文章编号:1671-6833(2010)01-0120-05

一种基于内容的文档图像检索方法

宋 涛[1], 刘 刚[1,2]

(1.河南大学 计算机与信息工程学院,河南 开封 475001;2.河南省招生办公室,河南 郑州 450002)

摘 要:使用一个图像作为查询检索输入,根据该图像的版面分析特征、统计特征、纹理特征与数据库中图像的相似程度检索图像.该检索方法首先利用数学形态学对文档图像进行段落分割和行分割,作为文档图像的版面结构特征;然后根据图像的统计特征包括字符数、统计数特征、纹理特征给出文档图像抽取算法;最后给出检索算法模型.实验结果表明,本算法具有较好的查准率和查全率,在基于内容的文档图像检索中具有应用价值.

关键词:基于内容的图像检索;版面分析;文档图像;数学形态学;图像分割;图像特征

中图分类号:TP183 **文献标识码**:A

0 引言

随着多媒体技术的迅猛发展,图像信息的采集、传输、存储、处理、理解与综合应用成为计算机学科中最重要的研究领域之一.图像数据的爆炸性增长使得对图像的管理和检索成为关键,笔者研究基于内容的文档图像检索技术,针对文档图像的特点,使用一个图像作为查询系统的输入,在图像数据库中检索相似的图像作为输出.

基于内容的图像检索(CBIR)是利用图像本身的信息,通常以图像特征(颜色、纹理、形状与结构布局等)的相似性为检索依据,根据每幅图像都有的可比较特征进行检索[1].检索算法的核心问题是图像的相似度度量和图像的特征描述,这两类算法相互影响,只有两者结合时,才会得到较好的检索效果[2-3].

文献[1]采用层次匹配树进行图像检索,文献[4-5]使用基于区域的匹配解决这一问题.这些算法都是基于图像特征的提取、表达以及相似性计算.从低层特征中提取图像的语义描述,并根据语义相似来查询图像可以直接表达对图像的视觉感知,也更符合人们的习惯和要求.目前语义检索仍处于探索阶段,主要研究都集中在简单语义如图像类别或基于类别知识的目标检索的研究.

笔者针对文档图像,图像的相似性定义为具有类似的版面结构(段落,行)、统计特征及纹理特性(粗糙度、对比度、方向度、粗略度等[9]).笔者先介绍文档图像及其特点,据此给出文档图像的特征抽取算法,最后给出检索算法模型和实验结果,并说明算法的有效性.

1 文档图像的特征抽取

1.1 文档图像特征

笔者所研究的对象为手写体扫描图像,图像中不包含有非文字信息.与非文档图像相比,文档图像的特点主要体现在以下几个方面:

(1)文档图像的直方图具有不均匀性,图像的信息熵小,在图像内容检索中有利于图像分类,有些算法直接利用直方图对图像进行检索,并作为图像数据库检索系统中重要的研究内容.

(2)有效的文字信息对应的像素所占比重少,但分布于图像的各个区域,当分散性较强时,不利于压缩,而且仅需要比较少的灰度级就可以表示文档图像,因此对二值文档图像的压缩是一个重要的研究应用领域.

(3)图像边缘丰富,图像的能量主要集中在低频区域,均值、方差、能量大.

(4)字符特征显著,文档图像的粘连字符分割、自动分类、识别算法复杂,其中手写体汉字识别是公认的世界级难题.

收稿日期:2009-08-18;**修订日期**:2009-10-29

作者简介:宋涛(1982-),男,河南驻马店人,硕士研究生,研究方向为数据挖掘.**通信作者**:刘刚(1962-),男,河南信阳人,博士,研究方向为数据挖掘.

图 1 为文档图像与 Lena 图像的直方图特性,(a),(b)为两个原始图像,(c),(d)是两个图像的一维直方图,(e),(f)为两个图像的二维直方图,从图中可以看出,一般图像的一维直方图连续性较强,而文档图像离散性强,二维直方图在计算时,采用了邻域(5 * 5)内的平均值,反映在二维直方图的图像上,能量主要集中在对角线,说明一般图像具有低频特性.

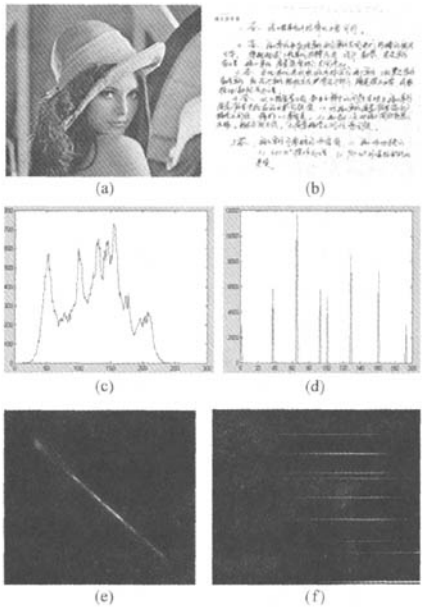

图 1　文档图像与 lena 图像的直方图特性

Fig. 1　Histogram features of document image and lena image

1.2　文档图像的版面结构分析

笔者使用数学形态学方法对文档图像进行段落标记,得到段落子图,然后对子图像进行行分割.为有效对空格信息进行标记,需要预估计文档图像中有效文字的宽度及文字之间的间距,段落空格的宽度是两行文字之间的间距与一个文字宽度之和,这是有效区分段落的基本特征.

如图 2 所示,两条实线之间表示行块分割,段落块的特征用两个行间距和一个行宽度表示,即两条虚线之间.

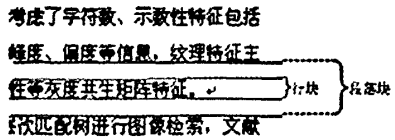

图 2　行块与段落块的表示

Fig. 2　Indications of line and paragraph blocks

利用形态学腐蚀算子对文档图像做膨胀运算,膨胀模板记为:$f(d,l)$,其中 d 为段落空格的宽度,l 为有效段落空格长度.顺序扫描整个图像,完成文档图像的结构分析.

如图 3 所示,文档图像经过版面分析后,获取到其结构框架,普通线的线代表行信息,黑粗线线代表段落信息,之后使用霍夫变换计算蓝色线及红色线的长度、倾斜角,从而完成了图像向文本级数据(图像语义)转换.

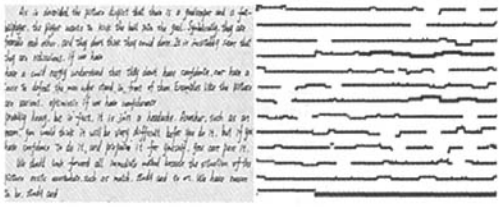

图 3　行和段落特征描述效果

Fig. 3　Description results of line and paragraph features

图 4 为版面分析算法的主要流程图.其中文本区域的检测使用投影算法,使用最大熵分割算法将图像的前景和背景分离,为行块、段落块的标记算法作预处理,(d,l)的估计非常重要,对版面分析影响较大,也是本算法进行实用化的基础(手写体文档字符大小,行间距多变),笔者使用一种较小的标记算法解决这一问题,细化可以使 Hough 变换的效率大大提高.由于文本图像的段落标记具有水平直线特点,通过实验观察,倾斜角的范围可定义为区间(-10,10)角度,Hough 变换的参数平面分别反映了行、段落标记的长度.

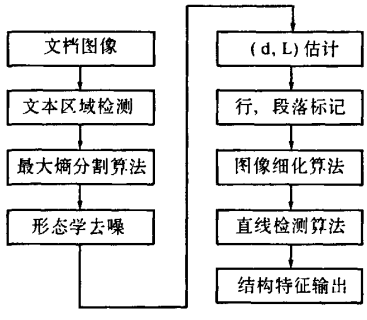

图 4　版面分析算法流程图

Fig. 4　The flow chart of layout algorithm analysis

如图 5 所示,(d,l)的估计算法的主要流程图,初始化模版参数是固定的,笔者采用的 7 × 7 的模版,目的是标记图像中的行文字,细化是将保留行的基本特征并有利于倾斜角检测的精度,倾斜角检测使用 Hough 变换,图像旋转后采用双三

次插值算法就行纠偏,双三次插值算法是双线性插值算法的改进算法,具有更好的纠偏效果,算法的时间复杂度较高.

图6为倾斜角检测过程示意图,左边为原始图像,中间是标记结果,右边是细化后的结果,检测精度可以到0.1角度.

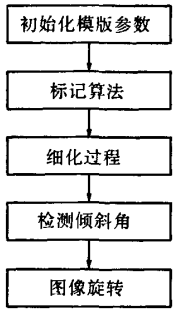

图5 (d,l)的估计算法流程

Fig.5 Estimation algorithm of d, l

图6 倾斜角检测过程示意图

Fig.6 Diagram of the process of sliding angle test

针对文档图像的检索定义为查询图像与数据库中的图像具有相似的版面及其统计特征,结构定义为段落和行,结构的特征定义为段落空格的长度,文字行的长度,图像的倾斜角.

笔者对文档图像的结构分解采用树型分解,段落和行特征信息分别存储,在检索策略上方便用户的使用,例如有些情况下根据段落检索图像,有时候需要按行特征进行检索图像.

经上述定义,文档图像的版面描述可用段落特征点集P和行块特征点集L表述.

$$P=\{p_i \mid i=1,2,\cdots,m\}$$

$$L=\{l_{i,j} \mid i=1,2,\cdots,m,j=1,2,\cdots,n\}$$

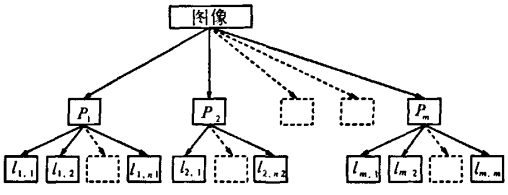

图7 图像版面树型分解

Fig.7 Tree structure for image layout

如图7所示,文档图像的版面分析示意图,上层是图像数据,中间层是段落标记,其值的含义为每个段落所包含的文本行数,最下面一层给出详细信息,每个行包含的文本行长度信息和统计示数性特征.为了提高检索速度,在抽取的特征上建立位图索引.

1.3 纹理特征分析

可从灰度共生矩阵比较他们的纹理特性,常用的基于灰度共生矩阵统计特征有以下几个:

(1)能量:反映了图像的尺寸及暗像素所占比重.

$$P=\sum_i\sum_j p(i,j)^2$$

(2)信息熵:反映了图像所含信息熵的多少.

$$E=-\sum_i\sum_j p(i,j)*\log p(i,j)$$

(3)对比度:反映背景与前景的差异程度,表征图像的清晰度的重要指标.

$$C=\sum_i\sum_j (i-j)^2*p(i,j)$$

(4)一致性:反映图像相邻像素灰度级的连续性.

$$U=\sum_i\sum_j \frac{P(i,j)}{1+|i-j|}$$

另外,文档图像的清晰度可作为宏观纹理特征进行分析,并作为文档图像的全局特征.清晰度通过一维直方图进行表达,实验表明清晰度能恰当的分类文档图像,有利于提高检索结果的精度.

如图8所示,根据主观评价评选出20幅清晰图像,20幅模糊图像,计算他们之间的相关性,蓝色清晰图像之间的标准相关都在0.85以上,模糊图像之间的相关性大都落在(0.3~0.85)之间,清晰图像和模糊图像之间的相关性比较差,大都在0.4以下.

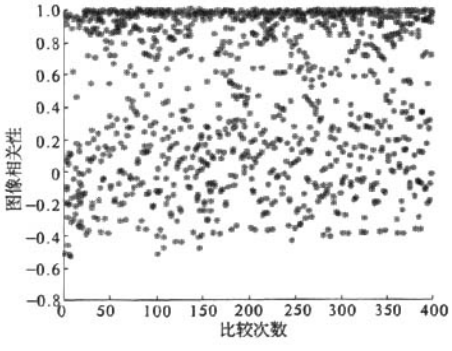

图8 图像直方图之间的相关性

Fig.8 Correlations of image histograms

2　检索算法设计

2.1　基于内容的图像检索系统结构

笔者针对文档图像检索问题,数据库中包含有对图像特征的描述,检索针对的数据通常是从原始图像中抽取出来的高维特征向量[10].

CBIR 系统主要包含图像特征抽取与表示、多维索引技术和检索引擎,高维索引结构是 CBIR 系统必不可少的组成部分,特别是对于大规模或超大规模图像数据库尤为如此.

图 9 为检索系统框图. 检索结构是实用化过程中的核心技术,在大量的文档图像抽取的特征上建立合适的索引结构是必须的,多种检索结构相结合以充分提高检索效率,例如段落可以采用位图索引,对于纹理、清晰度等建立 B + 树索引,以充分提高应用灵活性,例如范围查询.

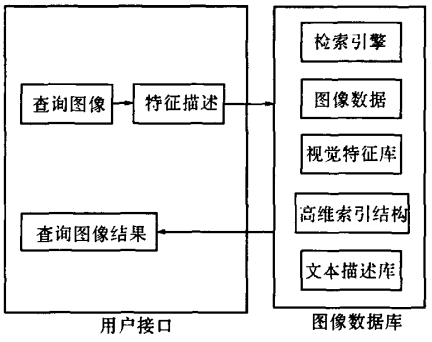

图 9　检索系统结构

Fig. 9　Retrieval system structure

2.2　检索算法实现

实现图像的检索主要由三部分构成:

(1)版面分析、统计特征、纹理特征计算.

(2)建立特征库.

(3)图像特征匹配算法实现,返回图像检索结果.

下面给出检索实现的步骤:

(1)输入查询图像,进行版面分析、特征计算.

(2)精确检索段落.

(3)模糊检索行的长度和特征,定义特征点集相似性,笔者将文档图像特征抽象成空间中的点,两个图像的相似性度量定义为他们之间的欧氏距离.

(4)二次检索,第一次检索使用行的长度,根据检索结果取前 n 个图像,如果能够满足用户的需要,则停止,返回图像检索结果,否则,进行特征匹配.

(5)取检索结果的前 m 个图像作为输出.

2.3　检索效果的评价

在基于内容的检索中,需要对算法的有效性进行评价,对检索效果评价主要使用的是查准率(precision)和查全率(recall)两个指标.

查全率的含义是在一次查询过程中,用户所查到的相关图像的数目和数据库中与目标图像相关的所有图像数目之比,查全率反映系统检索相关图像的能力. 查准率指在一次查询过程中所查到的相关图像数目同所有符合条件的图像数目之比,查准率反映系统拒绝无关图像的能力. 笔者使用这两个指标对检索的效果进行评价.

3　实验结果与分析

为了测试检索算法的有效性,进行相关的实验,数据为手写体汉字,分辨率 100dpi,灰度 256 级图像,数据量为 16 732.

为测试系统的检索效果,进行分组的相关实验如下:

①版面结构图像检索;②特征图像检索.

表 1 为多段落多行检索结果,可以指定段落数及每个段落包含的行数和段落空格的长度,长度以像素为单位,实验表明检索出的匹配图像数较少,其查准率为 1.

表 1　多段落多行检索结果

Tab. 1　Search results of multi - paragraphs and multi - lines

检索条件	段序号	每段行数	段长度/像素	检索结果	检索错误数	查准率/%
	1	2	50			
条件 1	2	5	50	7	0	100
	3	2	50			
	1	1	60			
条件 2	2	5	40	7	0	100
	3	3	50			

使用的统计特征包括:信息熵、均值、方差、偏度、峰度、能量;纹理特征包括:二维能量、二维熵、对比度、一致性、清晰度. 使用 B + 树在特征上建立索引结构,提高检索效率,并可提供范围查询,每个统计特征归一化到[0,1]区间,最后取其平均值.

如表 2 所示,统计特征检索的结果,范围选取 0.8 以上,属于高度相关,在返回的图像中包含有输入的检索图像,11 个统计特征具有独立性,检索的准确性是较高的.

表2 特征检索结果

Tab.2 Search results of features

图像编号	检索类别	范围	返回图像数	查准率/%
1	统计特征	[0,8,1]	3	100
2	统计特征	[0,8,1]	5	100

4 结束语

笔者研究了一种基于内容的文档图像检索方法,建立文档图像版面结构分析,使用高维特征向量描述文档图像特征,在检索策略上兼顾了精确匹配和模糊查询,检索时返回的图像数较少,精度较高.

检索算法实验表明:检索模型具有稳定性,版面分析能处理图像偏斜的情况,图像的特征抽取和计算简单。评价结果表明,该算法具有较高的查准率,具备一定的实用性.

文档图像的结构复杂,手写体文档图像更是如此,文档图像的分割、版面分析对实验结果影响很大,另外,需要对文档图像的子图检索做更深入的研究以及提出更加灵活和方便的索引结构.

参考文献:

[1] LUO J, MARIO A N. Content based sub - image retrieval via hierarchical tree matching[C]. ACM - MMDB,2003. 63 - 69.

[2] MANOUVIER M, RUKOZ M, JOMIER G. A generalized metric distance between hierarchically partitioned images[C]. MDM/KDD'05 August 21,2005, Chicago USA. 33 - 41.

[3] 丘衍航. 基于GMEM聚类的EMD图像检索[C]. 第十三届全国图象图形学学术会议, 2006. 575 - 579.

[4] CARSON C, THOMAS M, BELONGIE S, et al. A system for region - based image indexing and retrieval [C]. In Proc. of 3rd Intl. Conf. onVisual Information Systems, 1999. 509 - 516.

[5] LI J, WANG J Z, WIEDERHOLD G. IRM: Intergrated Region Matching for Image Retrieval[C]. In Proc. of ACM Intl. Conf. on Multimedia, 2000. 147 - 156.

[6] VALTTERI T, TIMO A, MATTI P. Block - based methods for image retrieval using local binary patterns [J]. SCIA 2005, 882 - 891.

[7] 金磊,陈优广,严敏. 一种基于用户感兴趣区域的图像检索算法[J]. 计算机技术与发展, 2006, 16(3).

[8] 刘涛,张艳宁,孙瑾秋. 一种基于目标区域的图像检索方法[J]. 计算机工程与应用, 2006, 42(26): 68 - 70.

[9] MA W Y, ZHANG HJ. Bench marking of image features for content - based retrieval [C]. California, USA: The Thirty - Second Asilomar Conference on Signals,1998. 253 - 257.

[10] 贺玲,吴玲达,蔡益朝. 一种面向大规模图像库的降维索引新方法[J]. 计算机工程, 2006,32(22): 20 - 22.

[11] SMEULDERS A W M, GEUSEBROEK J M, GEVERS T. Invariant representation in image processing. IEEE international Conference on Image Processing[C]. IEEE Computer Society:2001,18 - 21.

A Content - Based Algorithm for Document Image Retrieval

SONG Tao[1], LIU Gang[1,2]

(1. Institute of Computer and Information Engineering, Henan University, Kaifeng 475001, China; 2. Higer Education Admission Office of Henan, Zhengzhou 45002, China)

Abstract: This paper studies the content - based image retrieval for document image. Given a query image, the system returns overall similar images by layout analysis and statistic feature in image database. First, segment an image into paragraphs and lines based on mathematical morphology, return the image layout analysis results; and then compute the image statistic feature include characters, statistic count feature and texture to give distil arithmetic of the document image. In the end, we describe the matching model. This algorithm is tested through trials and errors. The experiment results indicate this algorithm is good at precision and recall. This algorithm is highly valuable in document image retrieval.

Key words: CBIR; layout analysis; document image; mathematical morphology; image segmentation; image feature