基于投影直方图的文档图像快速匹配研究

王 丹[1],刘 江[2]

(1. 山东师范大学 信息科学与工程学院,山东 济南 250014;
2. 山东山大鸥玛软件有限公司 数据研究中心,山东 济南 250100)

摘 要:如何实现文档图像间的快速匹配已成了人们日益关注的课题。针对文档图像的现有方法进行了研究,提出一种文档图像匹配的新方法。为进一步提高文档图像的匹配性能,结合用行列方向投影特征进行了文档图像的特征分析和提取工作,从而建立了特征相似匹配模型,在进行文档图像匹配时主要采用了平方差进行相似度度量和折半查找进行快速匹配的策略。实验表明随着二分法次数的增加,文档图像的相似度比较效率一定程度上得到很大的提高,匹配算法具有很好的抗倾斜和抗压缩效果。

关键词:相似度度量;投影直方图;平方差;折半查找

中图分类号:TN911.73 **文献标识码**:A **文章编号**:1673-629X(2011)07-0129-03

Study on Document Image Fast Matching Based on Projection-histogram

WANG Dan[1], LIU Jiang[2]

(1. School of Information Science & Engineering, Shandong Normal University, Jinan 250014, China;
2. Data Research Center, Shandong Shandaoma Software Co. Ltd, Jinan 250100, China)

Abstract: How to realize the rapid matching among document images has become the people increasingly focus on the subject. Presents a new method of document image matching aiming at existing methods. In order to improve the matching performance of document images, paragraph features and row features are used to extract feature and build similar matching model in the paper. Use square error(SE) and binary search when similarity of document images are matched. The experiment shows that with the increase of dichotomy numbers, the efficiency distinctly of document images is greatly improved to some extent. Thus the algorithm is good at precision and robustness.

Key words: similarity measurement; projection histogram; square error; binary search

0 引 言

随着图像采集、存储技术和Internet的发展,越来越多的文档信息以文档图像形式保存和应用。文档图像检索被广泛应用于一些机构组织,如办公自动化、数字图书馆等领域[1]。在文档图像的检索中,文档图像的相似度度量已成为一项非常关键的技术。文档图像的相似度度量是根据纹理、版式、文字信息等进行形状、特征或数据比较相似性得到的一个度量函数,给出了文档图像间的相似程度量化描述,在图像匹配、信息检索、图像融合领域等都有着重要的应用,当面对海量级的文档图像库检索时,如何实现文档图像的快速匹配已成为一个十分关键的问题。

理论上文档图像可以通过OCR(Optical Character Recognition)转化为文本数据,然后进行快速的相似度计算。虽然目前OCR技术能够提供高的识别正确率,然而OCR耗时较多,而且遇到字符粘连或者变形时往往需要人的交互性来完成识别过程,而且各种识别技术的识别范围受到限制,如藏文等少数民族语言文字的识别尚待解决,识别局限于部分文种[2]。对于大规模的文档图像数据库来说使用OCR方式代价是非常大的,极大地限制了文档图像检索领域的应用。

近年来,许多研究者已经提出了关于文档图像匹配的算法。赵珊等引入了字符串匹配技术进行图像的相似度度量[3]。孙远等提出了KMP快速字符匹配算法对图像快速匹配[4]。罗钟铉等提出的灰度图像匹配的快速算法[5]。P. Herrmann等提出的基于页面几何特征文档图像匹配算法[6]。阳方林等引入了一种新的快速图像匹配方法,即通过计算图像的迹差完成图像的近似匹配[7]。

收稿日期:2010-12-24;**修回日期**:2011-03-03

基金项目:2008年度山东省中青年科学家科研奖励基金资助项目(2008BSB38001)

作者简介:王 丹(1987-),女,硕士研究生,研究方向为数字图像处理、图像检索。

这些方法有的是针对局部特征的，如字符度特征，容易受分辨率、噪声、扫描品质等影响；基于版面特征的文档图像匹配算法虽然是基于全局特征的，但对图像的分辨率能力较低，这都会降低文档图像的检索精度。文中结合文档图像的行列方向投影特征，提出了一类文档图像快速匹配算法，适用于中小型文档图像库中进行图像特征的相似度快速计算，建立的相似度模型具有一定的鲁棒性。

1 特征提取

1.1 文档图像特征分析

文档图像与一般的图像有区别又有相似特征，主要有以下特点：

(1)文档图像检索是以文字和图表为主体内容的一类特殊图像，在文档图像特征提取时根据文本特征和图表特征是进行文档相似度度量的关键。

(2)文档图像的内容间存在空隙，直方图离散型较大，边缘信息丰富，图像的能量一般集中在低频区域，其信息熵比较小。

(3)文档图像相似度度量的主要目的是确定图像数据库中有无输入图像复本，这对于网上阅卷作弊监督行为有至关重要的意义。

(4)字符特征显著。对于文档图像的文本区域行与行间与字符间都存在空隙，这有利于实现前景色与背景色的分离，即所谓的文档图像的二值化。

文中主要针对扫描文本段落图像，既可以是印刷体，也可以是手写体，文档中不包含非文本信息，结合文档图像的行列方向投影特征，主要针对文档图像有效区域的段落和行分别进行特征提取，建立相似模型来实现文档图像特征的快速匹配工作。

1.2 文档图像特征提取

针对文档图像的特点，基于投影函数和数学形态学进行段落检测[8]，所涉及的特征可包括段落特征(段落数，段落长宽比)，行特征(行数、行长度、行高度、行间距)。

为了实现文档图像灰度一致性，首先对文档图像进行二值化处理。考虑图像空间分布特征，采用了投影直方图方法对文档图像有效区域进行定位。文档图像各像素点的值分别在水平和垂直方向上进行投影[8,9]，这样就可以得到文档图像的有效区域。对有效区域进行水平和垂直方向上光栅扫描，计算前景色像素值的个数，得到水平直方图 H 和垂直直方图 V，通过投影直方图可以得出有效区域的长 h 和宽 w，有效区域的长宽比定义为：

$P2=\frac{h}{w}$，所求得的特征向量为 $H1\{P1,P2\}$。

对图 1 的文档图像进行行列方向投影可以得到图 2 水平投影直方图和图 3 垂直投影直方图。通过上述的图示可以得出文本行的相关信息。

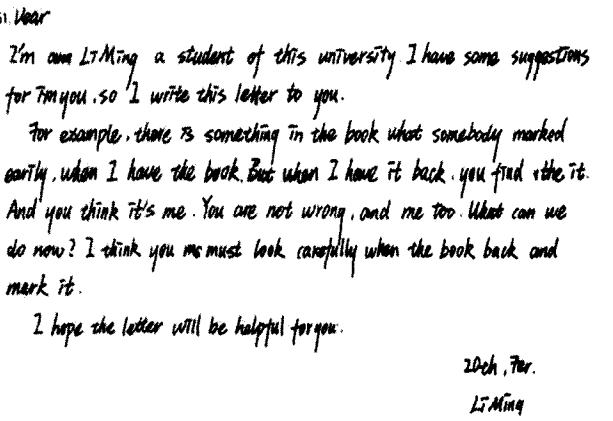

图 1 原始文档图像

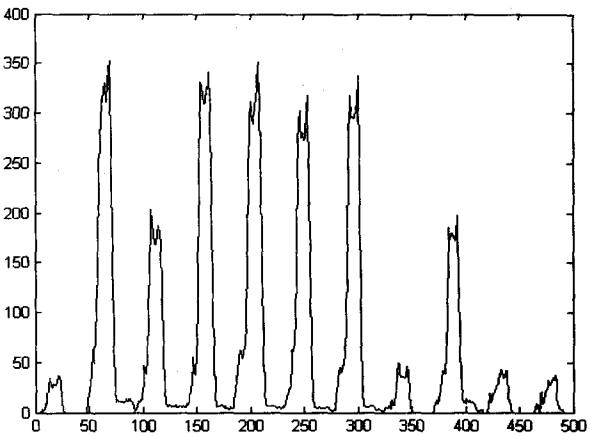

图 2 文档图像水平投影直方图

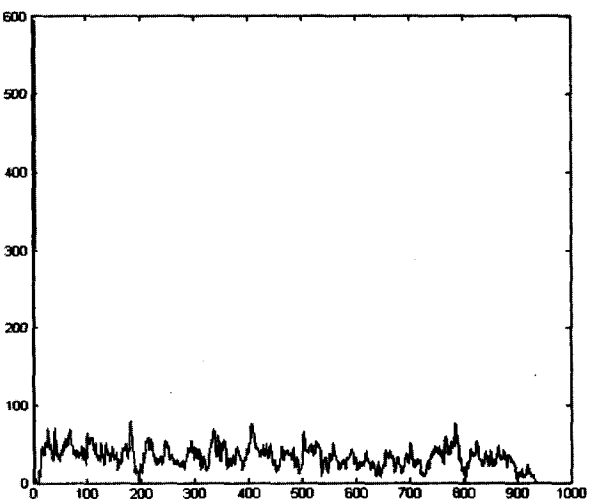

图 3 文档图像垂直投影直方图

针对文档图像的特点，使用数学形态学方法中的膨胀运算可以对文档图像进行段落分割，可求得段落数目，记为 $R1$。然后对段落再进行行分割，由于一般每行在水平方向的投影必然在黑像素间取得局部小，因此字符的分割点可设定在该最小值对应的坐标的小邻域内，对图像进行水平方向投影生成直方图求得行数 $R1$。同理，对每行的字符从左到右进行扫描进行垂

直方向投影可得到行长度 $R2$,所求得的特征向量:$H2\{R1,R2\}$。

由于段落特征和行特征在文档图像发生局部位移、旋转、缩放时是不随着发生改变的,具有一定的鲁棒性和适应性。这样就建立了相似模型:段落特征:段落数,段落长宽比;行特征:行数,行长度,行高度,行间距。通过两个特征的相似度比较可进行文档图像的近似匹配。

2　相似度度量模型

文档图像的相似性度量方法有好多种,大体可以归为两大类:距离度量和相关度量[9]。距离度量可以达到文档图像近似匹配的目的,两幅文档图像越相似,则二者之间的距离越小,常见的方法为绝对差度量、方差度量等。相关度量是通过比较两幅文档图像特征矢量的内积或其之间夹角的余弦值来定义的,常用的方法为积相关、相位相关算法等,在图像信噪比降低时方差度量法优于绝对差度量法[10]。积相关算法[11]在理想的情况下,其度量的极大值不一定是唯一的,容易导致匹配错误。文中采用距离度量法中的方差度量进行文档图像的特征相似度比较。

采取一种先粗后精的分层策略有效地对文档图像进行相似度度量,首先对特征提取后的文档图像特征库利用方差度量法分别进行段落特征比较和行特征比较,并进行 K 均值分类,然后进行折半查找选出与指定文档图像的特征近似匹配的候选集,此时完成了文档图像特征的粗匹配,然后查询图像与候选集中的元素进行逐个比较,得到最终的候选集,这样也大大地提高了特征相似度比较效率。

(1) 对于指定的图像 q,计算段落特征和行特征,分别记为 $H1(q)$,$H2(q)$;类似对图像库中的任意一幅图像计算特征:段落特征和行特征,建立特征库。

(2) 对文档图像库中的各个图像 $P(t)$ 分别进行段落特征与行特征之间的比较,计算公式为

$$\mathrm{Dis}_1(i)=\sum_{j=1}^{2}(P_j(i)-P_j(t))^2 \tag{1}$$

$$\mathrm{Dis}_2(i)=\sum_{j=1}^{2}(R_j(i)-R_j(t))^2 \tag{2}$$

其中 $i,t\in N$　N 图像库中的图像数。

求出图像库中特征差异的平均值:

$$A_1^1=\frac{\sum_{i=1}^{N}\mathrm{Dis}_1(i)}{N},A_2^1=\frac{\sum_{i=1}^{N}\mathrm{Dis}_2(i)}{N},$$

经过 k 次均值分类可得到文档集:

$$\mathrm{Set}_1^k=\{(\mathrm{Dis}_1(i)\le A_1^k)\cup(\mathrm{Dis}_2(i)\le B_1^k)\} \tag{3}$$

$$\mathrm{Set}_2^k=\{(\mathrm{Dis}_1(i)>A_2^k)\cup(\mathrm{Dis}_2(i)>B_2^k)\} \tag{4}$$

其中 $\mathrm{Set}_1^k\subset\mathrm{Set}_1^{k-1}$,$\mathrm{Set}_2^k\subset\mathrm{Set}_2^{k-1}$,这样把图像特征库划分成 2^k 个文档集,图像库中较为相似的图像特征都聚在同一集合中,这样就建立了特征库索引。

(3) 根据所建立的分类特征库索引,对(1)所计算出的指定图像 q 的段落和行特征在图像特征库中进行折半查找。找出最为相似的文档类集合,记为 S_1^m,S_2^n,选候选集为 $S=S_1^m\cup S_2^n$。

(4)对候选集中的图像进行排序。设定初始化权重 w_1 和 w_2,按照公式(1)(2) 计算集合 S 中图像的特征和指定图像 q 的特征差异,并且得到 $\mathrm{Dis}_1(i)$ 和 $\mathrm{Dis}_2(i)$,则文档图像的特征差异定义为

$$D(i)=w_1\mathrm{Dis}_1(i)+w_2\mathrm{Dis}_2(i) \tag{5}$$

文档图像的特征相似度定义为

$$\mathrm{Sim}(i)=1-D(i) \tag{6}$$

选取文档特征相似度大于相似度阈值 T 的文档图像,相似度按照递减顺序进行排列得到最终的匹配候选集。

3　实验结果分析

为了验证相似度度量方法对小型图像库是否有效,对山大欧码图像处理研究中心提供的 200 幅不同形式的扫描文档图像(灰度,彩色)进行了仿真实验。实验的环境为 Windows XP 操作系统,512M 内存,利用 Matlab7.0 建立了图像特征库,实现了文档图像的相似度度量,比较顺序查找和不同的 K 均值分类实现文档图像的相似度的平均查找时间比较。

实验中设定 $m=12$,初始权重 w 均为 0.5,K 分别为 2、4、6、8、10,所产生的实验结果如表 1,2 所示。

表 1　在 200 幅图像库中相似度阈值 T 为 0.5 的测试结果

算法	平均比较时间/s
顺序查找	10.90
K=1	8.91
K=3	5.55
K=5	2.40

表 2　在 200 幅图像库上相似度阈值 T 为 0.3 的测试结果

算法	平均比较时间/s
顺序查找	19.91
K=1	17.91
K=3	11.55
K=5	6.40

通过仿真实验可以看出,在文档图像特征的相似度比较方面,利用 K 均值分类比顺序查找时间复杂度

(下转第 137 页)

3 结束语

文中以引导学生快速初步掌握网络编程方法为目的,简明扼要地阐述了三种基本网络传送模式单播、组播和广播的编程实现思路和要点以及多线程的实现框架。此外,初步探讨了结合网络监听工具的网络程序调试方法和程序实际数据传送情况分析方法。

目前已经在课程设计中尝试先向学生介绍采用基本网络传送模式的应用程序的编写,然后再鼓励学生发挥主观能动性借鉴文中的数据传送分析和调试方法自己编写稍稍复杂的小程序,例如基于TCP的大文件传输工具、基于UDP的多方文字聊天工具等,发现不少同学已经不再惧怕网络编程,能够较好地完成课程设计任务。

参考文献:

[1] 王西锋,张晓李. 网络编程能力培养模式的探索与实践[J]. 计算机教育, 2009(2):93-94.

[2] 成卫青,杨哲睿. 网络编程实验设计与教学研究[J]. 实验科学与技术, 2010, 8(2):99-101.

[3] 刘　淼,刘怀亮. 计算机专业《网络编程》实验教学改革探索[J]. 实验室科学,2007(2):25-27.

[4] 刘　琰,常　斌,罗军勇,等. 面向能力培养的网络编程技术课程教学方法探讨[J]. 计算机教育, 2010(18):52-55.

[5] 李向丽,李　磊,陈　静. 网络实验仿真与网络技术实践[J]. 计算机技术与发展, 2006, 16(3):74-76.

[6] 李　鹏. 网络编程技术课程的教学改革思路[J]. 西安邮电学院学报, 2010, 15(2):166-168.

[7] 张晓明,杜天苍,秦彩云. 计算机网络编程课程的教学改革与实践[J]. 实验技术与管理, 2010, 27(2):4-7.

[8] Microsoft. socket Function[EB/OL]. 2011-01-20. http://msdn. microsoft. com/en-us/library/ms740506(v=vs. 85). aspx.

[9] Malhotra A, Sharma V, Gandhi P, et al. UDP based chat application[C]//2nd International Conference on Computer Engineering and Technology (ICCET). [s. l.]:[s. n.], 2010: 374-377.

[10] Law K L E, Leung R. A design and implementation of active network socket programming [C]//Eleventh International Conference on Computer Communications and Networks. [s. l.]:[s. n.], 2002: 78-83.

[11] Juszkiewicz K. UNIX Network Programming, Volume 1: The Sockets Networking [J]. IEEE Communications Magazine, 2004,42(5):20-21.

[12] 史蒂文,科　默. 用TCP/IP进行网际互连:设计、实现与内核(ANSI C版)(第2卷)[M]. 第3版. 张　娟,王　海,黄述真,译. 北京:电子工业出版社, 2008.

(上接第131页)

要低,在小型文档图像库中,根据特征对文档图像库进行分类,建立特征库索引,可以大大提高查找效率,实现文档图像间特征相似度的快速比较。

4 结　论

结合了行列方向投影特征对文档图像进行了相似度比较,其模型具有稳定性,当文档图像比例发生缩放、倾斜都有一定的鲁棒性。实验表明,在不同的K均值分类次数的情况下,综合利用段落特征和行特征进行K均值分类查找比顺序查找进行特征比较其效果要好,且随着迭代次数的增加,文档图像特征比较效率得到了提高。相似度度量的研究对下一步的文档图像检索提供了重要的参考依据。如何针对非文本文档图像的特征提取和相似度比较以及如何对中、大型文档图像特征库进行文档图像间的快速比较实现快速检索,也需要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] Liu Hong, Feng Suoqian, Zha Hongbin, et al. Document Image Retrieval Based on Density Distribution Feature and Key Block Feature[C]//Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition. [s. l.]: IEEE, 2005.

[2] 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 2006,46(7):1231-1234.

[3] 赵　珊, 郑清洁. 基于字符串匹配技术的图像检索算法[J]. 高技术通讯 2010,20(2):117-120.

[4] 孙　远, 周刚慧, 赵立初, 等. 灰度图像匹配的快速算法[J]. 上海交通大学学报,2000,34(5):702-704.

[5] 罗钟铉, 刘成明. 灰度图像匹配的快速算法[J]. 计算机辅助设计与图形学学报,2005,17(5): 967-969.

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

[7] 阳方林. 一种新的快速图像匹配算法[J]. 计算机工程与应用,2005(5):51-52.

[8] 王佐林,王希常,刘　江. 基于数学形态学的文档图像段落标记及其应用[J]. 山东师范大学学报(自然科学版), 2007,22(4):27 - 29.

[9] 张　田,王希常,陈昌华. 基于特征的文档图像检索[J]. 计算机工程,2009,35(22):176-178.

[10] 刘宝生,闫莉萍. 几种经典相似性度量的比较研究[J]. 计算机应用研究,2006(11):1-3.

[11] 陈卫兵. 几种图像相似性度量的匹配性能比较[J]. 计算机应用,2010,30(1): 98-99