一种基于版面结构距离的文档图像检索算法

赵 慧[1],王希常[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:介绍了一种基于版面结构距离的文档图像检索算法,使用版面特征作为文档图像的特征检索图像。先将文档图像进行梯度和最大梯度差(MGD)计算,然后使用MGD值作为一个窗口对文本区域进行融合,将文档图像以行线的形式标示出来。同时给出了检索的匹配方法,并对匹配方法进行了实验。实验结果表明,该检索方法具有较高的查准率,具有很好的抗倾斜和抗缩放效果。

关键词:文档图像;版面分析;文档图像检索;图像匹配
中图分类号:TP751 **文献标识码**:A **文章编号**:1674-7720(2010)21-0042-03

Document image retrieval method based on layout distance measures

ZHAO Hui[1], WANG Xi Chang[1], LIU Jiang[2]
(1.Department of Information Science and Engineering, Shandong Normal University, Jinan 250014, China;
2.Research Institute of Data Processing Oumasoft, Shandong University, Jinan 250100, China)

Abstract: This paper presented a method for document image retrieval by layout analysis as a search feature. At first, compute the gradient and the maximum gradient different (MGD)for the input document image, then the MGD values are filled within a local window to merge text segment. The line of document image is marked in the form of line. This paper gives the mach algorithm between the two images, tests the algorithm by image skewing and scaling. The experiment indicates that the algorithm is good at precision and robustness.

Key words: document image; layout analysis; document image retrieval; image matching

文档图像一般意为含有文字信息的图像,目前大多数信息是以数字化形式存在的,并以文档的形式组织起来存放在数据库中。在这样的数据库中查找有关资料其技术是关键。常见的文档图像检索方法是基于内容的文档图像检索(CBIR)。它是利用图像本身的信息,通常以图像特征(颜色、纹理、形状、结构布局和语义特征等)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。

虽然目前OCR技术已经能够提供很高的打印体字符识别正确率,但是往往需要人工交互来提高字符识别的正确性。这对一个大规模的文档图像数据库来说,其代价是相当大的。手写体字符的识别本身相当困难,而语言相关性也是这类算法的一个明显的缺点。因为不同的语言文字要求依靠不同的OCR系统去处理混合多种语言的文档图像,这将影响检索系统的使用范围。

BEUSEKOM J V. 等人提出了一种基于版面分析的文档图像检索的距离度量方法,将文本区域分为不同的矩形块,然后找到块的中心点,利用角点的曼哈顿距离来计算块之间的距离,再利用三种不同的方法进行匹配[1];WONG K Y.使用游程平滑算法进行版面信息提取的方法[2];BREUEL T M.提出了使用Whitespace算法来提取版面信息[3]。

本文提出了一种在文档图像数据库中使用版面特征进行检索的方法,具体定义了文档页面中均具有的行的版面特征。该方法直接作用于图像数据,具有抗倾斜和抗缩放的好处。

具体步骤是先将文档图像进行梯度和最大梯度差MGD(Maximum Gradient Different)计算[4],然后使用MGD值作为一个窗口对文本区域进行融合,提取出行块并用行线的形式标示出来,计算出相对坐标,再计算两版面

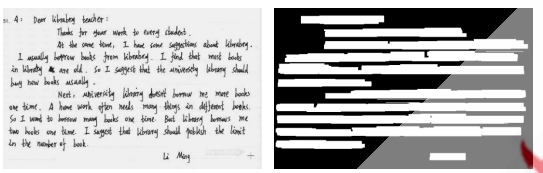
之间的距离进行匹配。

1 相关工作

1.1 文本行标记

将得到的文档图像进行预处理，具体的处理方法是：

使用文本行标记算法实现文字区域的行定位。本文使用[-1,0,1]对图像进行处理计算其梯度，然后计算其MGD。MGD计算方法如下：在一个大小为n的窗口内，用它的最大梯度差来进行填充，以达到文本融合的目的。因为英文和中文的字符宽度不同，根据具体的情况选择n，大于字符间距即可。将计算出来的梯度求它的最大值和最小值，然后相减，即为最大梯度差。将得到的MGD图像使用最大类间方差方法[5](OTSU)求出阈值得到二值图像[2]。图1为使用上述方法对行块进行标记的图像。



(a)手写原图像 (b)提取出的行块

图1 使用上述方法对行块进行标记的图

1.2 消除阶跃跳变

对于手写体或者英文的文档，会出现字符高低不一、笔画不连续等情况。线特征产生的断点可采用形态学方法、凸凹点处理和噪声处理三种基本策略提高直线的连续性，然后采用阶梯插补算法来消除阶跃跳变，算法的复杂度相对较低。

1	2	3
4	5	6
7	8	9

图2 3×3模板

在像素级上进行处理是：当出现行阶跃跳变的情况时，使用如图2的模板来对其进行填充。因为文档图像的行块在4个方向上都有可能出现这种阶跃，所以采用一个3×3的模板，以位置5为中心点，如图3所示，4种情况都包含其中：1和4为非文本像素，对4进行填充；3和6为非文本像素，对6进行填充；4和7为非文本像素，对4进行填充；6和9为非文本像素，对6进行填充。如果填充之后依然有符合结构的像素，则继续填充，即把需要填充的区域都填充完整。填充前后的图像如图4所示。

1.3 行线标记

通过对得到的二值图像的行跳变的填补，文本行的变化相对比较平滑，这有利于行线的标记。本方法取每个文本行的下边缘来作为行线。因为背景区域为黑色，文字区域为白色，所以对文档图像进行扫描，从黑色区域进入白色区域时所遇到的第一个像素进行标记，这样就把每一行的行线标记出来了，所得到的行线是单像素的。这种方法的优点是可以抗倾斜。

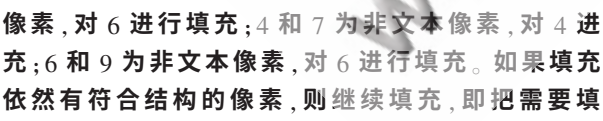
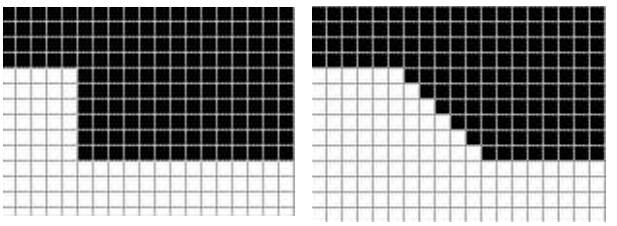


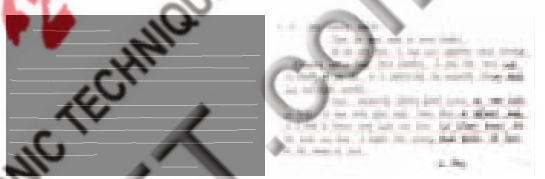
图3 阶跃跳变的四种可能情况



(a)填充前 (b)填充后

图4 填充前后的图像

图5(a)为对图1中的图像中的行用直线的方式标记出来。为了验证提取出的行线与原图是否一致，将它与原图(如图5(b)所示)进行了匹配，可以看出，所得结果是比较满意的。



(a)从图1中提取出的行线 (b)图1的原图

图5 对图1的图像行用直线标记出

2 匹配算法

本文所采用的方法是将行线抽象为空间中的一个点，点的灰度值定义为行线的长度。全局匹配模式考虑版面的加权平均，用于全局位置进行匹配，这个过程相当于文本区定位过程。局部匹配模式是定义两个行在位置、尺寸上的变化情况，通过位置优先(版面)得到匹配模式，进而对匹配误差能量进行计算。

匹配方法转化为两组点之间的匹配定义问题，点模式简化了问题的复杂性，只包含了版面结构信息、长度信息和尺寸信息。

(1)点模式匹配

假如一个页面上总共有m行，从第一行开始，因为行线为单像素，所以每一行的起始坐标为(x_i,y_i)，其中$i=0,\cdots,m-1$，将每行的长度定义为$z_i,i=0,\cdots,m-1$。总共有m个点，这m个点的中心点的位置坐标为(x_0,y_0)。在计算中心点的位置坐标时，将它的每行的长度作为权重考虑在内，即：

$$x_0=(\sum_{i=0}^{m-1}x_i\times z_i)/m,\ y_0=(\sum_{i=0}^{m-1}y_i\times z_i)/m$$

中心点加权匹配方式不能完全解决问题，图像在两个尺度上的缩放对这种方式影响极大。使用归一化的尺寸可部分解决这个问题，但归一化后仍需计算中心点的位置，通过中心点进行坐标转换，使用坐标转换后的新的点模式对差异性进行度量。

每一行起始坐标的相对坐标是 (x_i', y_i')，$x_i'=x_i-x_0$，$y_i'=y_i-y_0$。图 6 为将行线抽象为空间中的点的图像，其中亮度代表该行的长度，位置为起点坐标。

图 6 将行线抽象为空间中的点的图像

(2)距离匹配模式计算

将两个页面的中心点对齐，从第一个页面的第一行开始，与另一个页面每行进行比较。假如另一个页面的相对坐标是(u_j', v_j')，$j=0,\cdots,n-1$，每行长度为 w_j。计算两个待比较页面的坐标及长度的差 Δx_i、Δy_i、Δz_i，其中：$\Delta x_i=x_i'-u_j'$，$\Delta y_i=y_i'-v_j'$，$\Delta z_i=z_i-w_j$。则定义差异能量为：

$$d\mathit{Energe}(i)=\Delta x_i+\Delta y_i+\Delta z_i$$

将第一个页面的第一行与第二个页面的每一行进行比较，得到 n 个差异能量，求这 n 个差异能量的最小值 $\min(d\mathit{Energe}(i))$。第一个页面共有 m 行，将得到 m 个值，对其求和：

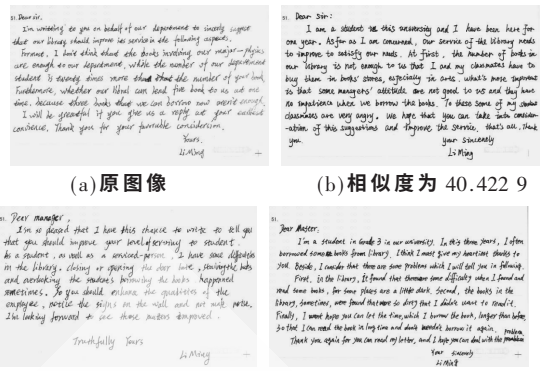
$$\mathrm{dis}=\sum_{i=0}^{m-1}\min(d\mathit{Energe}(i))$$

不匹配的情况经常发生，例如一个图像中含有 4 个点模式，另一个图像中含有 10 个点模式，内部点模式之间具有结构相关性，结构上的相关性定义为点模式位置掩模距离，该距离用来度量点模式全局匹配能力。如果一个点模式为另一个点模式的子模式，则该方法实现子图检索功能，模式距离最小时，产生最佳匹配。最佳匹配时，产生更为细致的行线检索能力。使用掩模方法是为了产生更好的查准率。

3 实验结果与分析

应用上述方法进行了实验，数据为手写体英文，数据采集分辨率为 100 dpi，256 级灰度图像，数据量为 100 幅文档图像。对不同的图像分别比较它们的相似度。图 7(b)、(c)、(d)是与图 7(a)的相似度分别为 40.422 9、45.760 7 和 43.407 8 的图像。图 8(b)、(c)、(d)是与图 8(a)原图像版面结构相似的几种图像类型。图 9(b)、(c)、(d)是与图 9(a)原图像版面结构具有差异的几种图像类型。

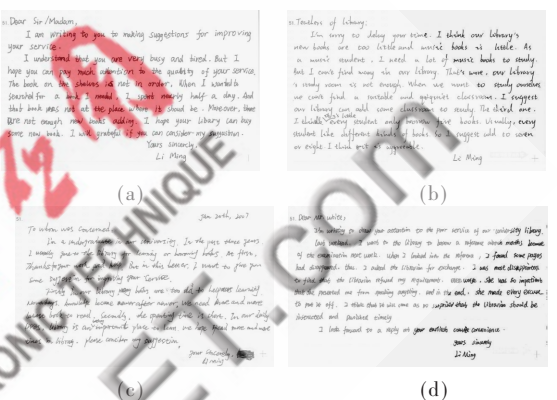
本文使用对 100 幅文档图像两两进行版面结构的匹配，共有 4 950 种结果。实验结果表明，两种不同版面的能量差异最大的在 340 左右，如图 10 所示。横坐标显示的是 100 幅图像两两匹配出现的情况的数目，可以取



(a)原图像 (b)相似度为 40.422 9

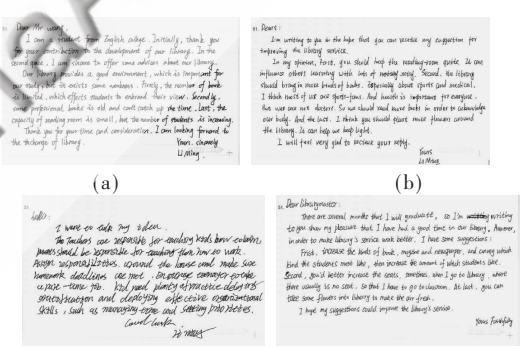
(c)相似度为 45.760 7 (d)相似度为 43.407 8

图 7 不同图像的比较



(a) (b) (c) (d)

图 8 与原图像(a)版面结构相似的图像类型



(a) (b) (c) (d)

图 9 与原图像(a)版面结构具有差异的图像类型

到的最大坐标为 4 950，纵坐标为各匹配情况对应的能量差异，最大值 350。从图中可以看出能量差异主要集中在 50~200 之间。

各个能量点的频数的直方图如图 11 所示，图中横坐标为能量差异数据，最大为 340 左右，提取到 350。纵坐标为取到各个能量的情况的数目的累加。从图 11 可以更直观地观察到能量差异在 50~200 之间的数目最多。

实验结果表明：(1)文档图像的版面结构具有相对的稳定性。(2)点匹配模式计算了最小距离，可有效表示图像的文本行基本信息。(3)距离匹配较为简单，使用了三个维度的一维距离，有较好的区分性。对距离计算统计

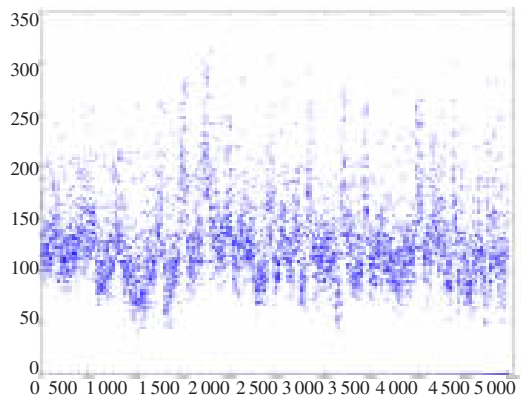


图 10 能量分布图

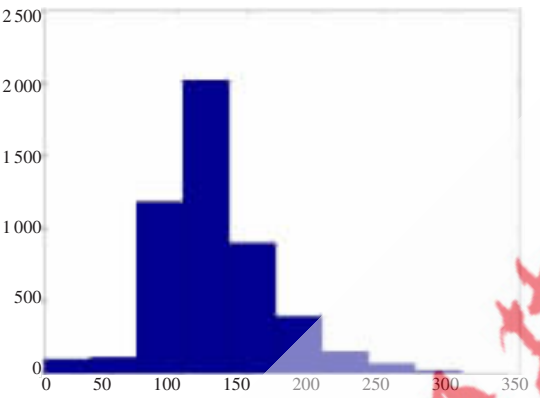


图 11 直立图频数

表明，具有正态分布特性。(4)点匹配模式需进一步进行研究，算法的复杂度需进一步降低，以进行实时图像处理。

本文针对文档图像的检索方法进行了研究，提出一种文档图像检索的新方法。分析了文档图像版面特性，使用分割方法确定文本行，将文本行进行标记，找出页面的中心点坐标，中心点坐标将文本行的长度作为权重考虑在内，得到相对坐标。根据相对坐标和文本行长度得到一个差异能量，根据差异能量来进行匹配。并对该方法进行了实验和结果分析。本方法的优点是，当文档的行出现倾斜和缩放时，不影响匹配的进行。但需要进一步降低所用的点匹配模式时间复杂度，以进行实时图像处理。

参考文献

[1] BEUSEKOM J V, KEYSERS D, SHAFAIT F, et al. Distance measures for layout-based document image retrieval[C]. In:2nd IEEE International Conference on Document Image Analysis for Libraries, yon, France, (2006): 232-242.

[2] WONG K Y, CASEY R G, WAHL F M. Document analysis system [J]. IBM Journal of Research and Development, 1982, 26(6): 647-656.

[3] BREUEL T M. Two geometric algorithms for layout analysis [C]. In DAS '02: Proceedings of the 5th International Workshop on Document Analysis Systems V, Springer-Verlag, London, UK, 2002: 188-199.

[4] JAE H K, TAE T P, YANG H C, et al. Photo-text segmentation in complex color document[C]. The 5th Japan-Korean Joint Symposium on Imaging Materials and Technologies, Kyoto, Japan, Nov. 2004: 44-47.

[5] OTSU N. A threshold selection method from gray-level histograms [J]. IEEE Trans. Systems, Man and Cybernetics, 1979, 9(1): 62-66.

(收稿日期：2010-07-05)

作者简介：

赵慧，女，1985 年生，硕士研究生，主要研究方向：图像处理。

王希常，男，1964 年生，研究员，博士，主要研究方向：图像处理。

刘江，男，1979 年生，硕士研究生，主要研究方向：图像处理。