基于视窗的 OCR 页面图像倾斜检测方法

靳 从 魏之来 杨静宇
(南京理工大学计算机系,南京 210094)

摘 要 文档在扫描输入过程中,所生成的页面图像一般都存在一定的角度倾斜,当页面图像倾斜角度过大时,将对进一步的版面分析以及字符识别产生不良影响。为了快速准确地检测页面图像倾斜角度和降低计算量,提出了一种基于视窗变换的页面图像倾斜检测方法,该算法首先对视窗中的文字及图片的细节部分进行模糊,然后对其边沿进行直线拟合,以便快速检测页面图像倾斜角度。实验结果表明,该方法能快速准确地检测出各类页面图像的倾斜角度,并具有良好的适应性。

关键词 图像处理 倾斜检测 文档图像
中图法分类号: TP391.41 **文献标识码**: A **文章编号**: 1006-8961(2004)11-1290-04

Skew Document Image Detection Method Based on Windows Transform

JIN Cong, WEI Zhi-lai, YANG Jing-yu
(Department of Computer, Nanjing University of Science and Technology, Nanjing 210094)

Abstract During OCR (optical character recognition) image scanning, the document images, are always placed slantwise to some extent. When the skew degree is big enough, it will influence the effect of document analysis and lower the recognition accuracy as the algorithm for layout analysis and character recognition are very sensitive to page skew. So the skew degree detection is a very important step during the preprocessing of document analysis. In this paper, a skew detection method based on the window analysis is presented. First it chooses the suitable windows which are not in the margin but in the layout of a printed page. Then according to the kind of contents, just like tables, text lines, images and etc., it uses the different methods to pre-processing the windows image. To overcome the large computing, the third step is to blur the text lines and image from the window. The forth step is to detect the edges of the blurring regions. At last it uses a straight line fitting to the edges, and gets the skew angle. By this method, experimental results show that the skew angles of many kinds of document images can be efficiently and accurately detected, and it has sufficient adaptability.

Keywords image processing, skew detection, document image

1 引 言

随着计算机性能的不断提高,各种情报资料的无纸化(即电子档案)需求越来越高。由于电子档案具有保存周期长、安全性高、占用空间少等优点,所以在计算机普遍使用的今天可广泛应用于各情报领域。电子档案的来源一般分为以下两类:一类是计算机广泛应用后手工录入的,即直接以电子格式文件存储的情报档案;另一类是在计算机大量使用前已形成的书面档案,其文字形式的文档一般可采用扫描仪等装置转换成数字格式,存储于介质,而扫描形成的页面图像则必须进行相应的处理,方能按可编辑格式加以存储。

页面图像处理包括版面分析、文本段落、行和单字的切分以及最后的字符识别[1]。在文字识别和表格自动录入软件中,页面图像的倾斜校正是一项重要的预处理技术[2],之所以重要,其基本原因在于:(1)页面图像的倾斜使 OCR (optical character recognition)中字符分割发生困难;(2)较大的倾斜会引起字符明显变形,致使大部分 OCR 方法难以适应;(3)在表格处理中,页面图像的倾斜会引起表格校

收稿日期:2003-09-04;改回日期:2004-05-10

正、识别和表格中固有信息的去除发生困难[3]。由于在文档图像扫描输入过程中,让用户在输入时保证无倾斜是很难做到的,因此在版面分析之前,对整个页面图像进行倾斜检测和校正十分必要。

目前,页面图像倾斜角度的自动检测方法主要分为基于投影图的方法、基于 Hough 变换的方法、基于交叉相关性的方法、基于 Fourier 变换的方法和基于 K-最近邻簇的方法等 5 类。

其中基于投影图的方法是利用投影图的某些特性来进行判断,而 Hough 变换是最常用的倾斜检测方法,由于基于 Hough 变换的 BHT 算法计算量非常大,所以不断有学者提出一些专门用于倾斜检测的 Hough 变换改进算法,其核心思想是减少转换的数据量。如果利用行程编码来表示灰度图像,以去除边缘及图片的影响,则可选取一定范围内的行程值行作为 Hough 转换的候选点[4]和可以将整幅图像子区域中候选对象最底部的像素点作为候选点,用于 Hough 转换[5]等。

本文所介绍的方法就是基于投影图分析的方法,同时借鉴了文献[5]中的子区域思路。

2　倾斜检测算法

由于无论页面图像包含什么内容,整个页面图像倾斜角度与局部页面图像的倾斜角度总是一致的,所以算法首先选择了一个视窗,并对视窗进行旋转,若找到窄幅脉冲,则认为视窗中存在直线,然后以该直线上的点为特征点进行直线拟合,即可得出页面图像的倾斜角度;若得到周期性脉冲,则认为所检测的角度小于 2°,此时应再通过对视窗图像内容进行模糊来提取相应的特征点,以便进一步利用直线拟合的方法来确定页面图像的倾斜角度。

2.1　视窗的选取

由于科技文档页面图像中的文本行一般采取横排,且相邻文本行之间的距离,即行距固定,因此其水平投影会呈现周期性。由于经过扫描转换的文档图像边缘通常会出现大段的噪声,所以在选择视窗时,要考虑去除整个文档图像边缘区域,否则将影响文档页面图像倾斜角度的检测,同时视窗应保证足够大才能判断文档页面图像倾斜角度,即视窗区域 K 中点的坐标应满足下式:

$$K_{x,y}=\{(x,y)\,|\,w_1\leqslant x\leqslant w_2,h_1\leqslant y\leqslant h_2,\ (h_2-h_1)\geqslant nh\} \quad (1)$$

设页面图像高度为 H,页面宽度为 W,视窗左边界 $w_1=W/8$,右边界 $w_2=W/4$,上边界 $h_1=H/6$,下边界 $h_2=H/3$;h 为行高,视窗包含的行数 $n\geqslant 3$。

若视窗中不包含图片、表格,而实际包含内容的文本行又小于 3,则视窗应重新选取。由于科技文档页面上主要包括文本、表格、图形或图片,因此在选取视窗后,通过水平、垂直两个方向的投影就可以进行视窗内容的检测(如图 1 所示)。

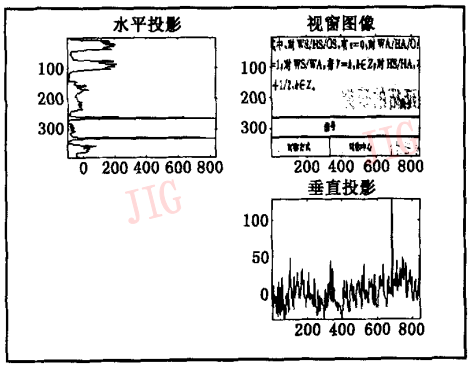

图 1　视窗中包含文本及直线

2.2　视窗的旋转

一般选取视窗中包含的图像内容不同,其投影图也不同。若选取的视窗中包含文本行,则对于小于 2°倾斜角的文档,可直接从其水平投影图中看到周期性脉冲,而且此类脉冲具有一定的宽度和峰值,若其中含有直线,则脉冲的峰值更大,幅宽更小(如图 1 所示,图中坐标原点为视窗左上角顶点,"水平投影"的纵坐标和"垂直投影"的横坐标分别为所选视窗的高和宽,单位为像素点数,而"水平投影"的横坐标和"垂直投影"的纵坐标的单位均为累计黑像素点,以下各图类同)。

若水平投影图中,脉冲不明显或脉冲的幅宽过小(如图 4 所示),则说明页面图像的倾斜角大于 2°或视窗中存在图片(如图 2 所示)。

算法提出将所选视窗图像进行旋转,以期得到相对最佳的周期脉冲(假设页面图像的倾斜角度范围在−45°～45°之间)。

采用视窗图像最左边为基线进行旋转,每一次旋转 2°,每旋转一次,均根据投影值进行周期性脉冲的判断,以便判断是否存在文本行(至少存在 3 个脉冲,即 3 个相对视窗的完整文本行)。当脉冲的峰值稳定后,即可得到近似的页面图像倾斜角度。

若存在一个窄幅脉冲,则可确定存在直线,可采用直线上的点作为特征点,转到"2.5 节页面图像倾

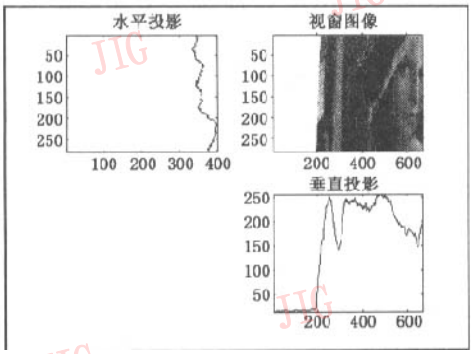

图 2 视窗中包含图片

斜角度的判断”，再通过直接进行直线拟合来得出页面图像的倾斜角。

2.3 二值化处理

若不存在直线，那么得到近似角度后，还需对视窗图像进行特征点的提取，以便进一步检测倾斜角度，即首先通过二值化处理去掉其中的噪声点。二值化是图像分割的一个重要部分，其通常是利用图像的主要提取目标与背景灰度特性上的差异来将目标与背景分割开来。图像二值化可根据下式进行阈值处理：

$$f_t(i,j)=\begin{cases}1 & f(i,j)\geqslant t\\ 0 & f(i,j)<t\end{cases} \tag{2}$$

二值图像中的值为 1 的部分表示目标图像，值为 0 的部分表示为背景。其中关键是阈值 t 的选取，经过多种阈值选取方法的试验，包括基于梯度均值的、基于熵的、基于矩的以及其他常用的方法，本文算法选定基于梯度均值的阈值选取方法，这是因为页面图像中背景与文字之间的灰度差异较为明显，且计算简单的缘故。具体实现方法为：

(1) 计算整个页面图像的梯度直方图；

(2) 将整个直方图进行归一化处理；

(3) 进行阈值 t 的计算；

2.4 模糊处理

因为文档页面图像的倾斜与页面的内容细节无关，所以算法可按下式将视窗内的图像进行水平模糊处理，以使得文字或图像部分的细节模糊。

$$f(x,y)=\begin{cases}1 & \sum_{i=\Delta d}^{R-\Delta d}\sum_{j=\Delta d}^{C-\Delta d}f(x_i\pm\Delta d,y_i\pm\Delta d)\geqslant\dfrac{d^2}{2},\ \Delta d=\dfrac{d}{2}\\ \text{不变} & \text{其他}\end{cases} \tag{3}$$

其中，R 为视窗水平方向像素总数，C 为视窗垂直方向像素总数，d 为字间距，式(3)表示对视窗图像中任意 $d\times d$ 大小的区域，若黑像素之和超过区域像素总和的一半，则将该区域像素值以黑像素替换，其结果如图 3 所示。

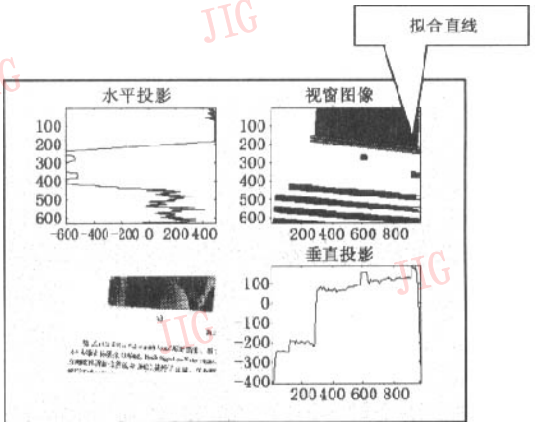

图 3 视窗中包含图片的处理结果

2.5 页面图像倾斜角度的判断

对于视窗图像，则由视窗最上沿开始，由上至下选取像素值发生变化最大的点为特征点，然后将相邻的特征点组成特征点集，若集合中的点数达到 100 个，则采用最小二乘法进行直线拟合。若整个视窗中符合条件的特征点不足，则从视窗最上沿开始，由左至右重新查找(查找图片的右边界)，注意此时检测出的角度应减去 90°。

若以图片的两条边或文本行开始的视窗，可直接通过上述方法来得到页面图像倾斜角度(如图 3、图 4 所示)。

若所选取的视窗中全部是图像，即不存在图像

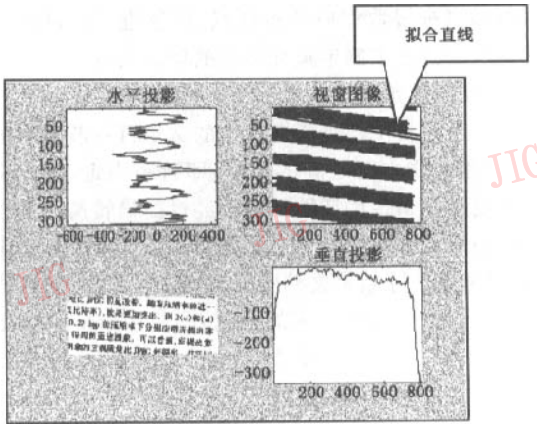

图 4 视窗中全文字的处理结果

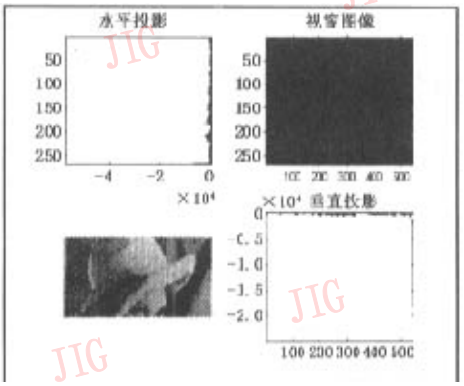

图 5　视窗中不包含图像边沿

边界(如图 5 所示),则选取以原视窗为右边界的同样大小视窗,通过再次进行模糊投影来进行页面图像倾斜角度的检测,否则重新进行视窗选取。

3　实验结果与分析

用于算法测试的页面图像为 A4 幅面大小的页面图像,它以 300 dpi 的分辨率扫描输入,内容包括中文文本、英文文本及带有图片、公式、表格等的页面图像。本文算法在奔腾 4 主频 1.7G 微机上实现,为了进行对比,本文将 Hough 算法及本算法分别应用于不同倾斜角的同一页面图像进行对比实验。本算法页面图像倾斜角检测的平均处理时间为 2.72 s,而 Hough 算法的平均处理时间为 20.3 s,表 1 列出对比实验结果及误差分析,其中标准偏差 σ 采用下式进行计算,其值越小,表示偏差越小,算法精确度越高。

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}\sigma_i^2}{n}} \tag{4}$$

表 1　两种算法实验结果对比

实际页面图像倾斜角度(°)	页面图像倾斜角度检测结果(°)	
	Hough 算法	本算法
−3.50	−3.62	−3.53
−2.50	−2.59	−2.51
−1.50	−1.60	−1.49
−0.50	−0.45	−0.47
−0.10	−0.07	−0.10
0.10	0.06	0.12
0.50	0.42	0.51
1.50	1.44	1.48
2.50	2.46	2.51
3.50	3.39	3.44
平均误差值	0.07	0.02
标准偏差	0.07	0.03

4　结　论

本文介绍了一个页面图像倾斜角度检测算法。为减少计算量,该算法仅选取整个文档图像的一个小区域,即视窗。检测时,通过对视窗图像进行旋转,并适当模糊,首先去掉其中的细节;然后选取像素值变化最大的点为特征点,通过直线拟合来计算出页面图像倾斜角。实验证明,该算法速度快,且准确度高。该方法可应用于文档分析的预处理,并可提高分析结果精度,但对于扫描页面图像质量过差,即噪声干扰太多时,由于视窗选取将极大地影响页面图像倾斜角度的检测结果,因此如何更好地选取视窗是下一步研究的方向。

参 考 文 献

1　Lu Y. Machine printed character segmentation-an overview[J]. Pattern Recognition, 1995, **28**(1): 67～80.

2　张忻中. 汉字识别技术[M]. 北京:清华大学出版社, 1992.

3　王姝华,李佐,蔡士杰. 基于最小二乘法的文档图像倾斜检测方法[J]. 计算机应用与软件, 2001, **18**(9): 43～46.

4　Le D S, Thoma G R, Wechsler H. Automation page orientation and skew angle detection for binary document image[J]. Pattern Recognition, 1997, **30**(10): 1325～1344.

5　Yu H, Jain A K. A robust and fast skew detection algorithm for generic documents[J]. Pattern Recognition, 1996, **29**(10): 1599～1629.

靳　从　1968 年生,副研究员,1995 年获南京理工大学计算机应用专业硕士学位。现主要研究领域为图像处理、中文信息处理、模式识别。
E-mail:j0805481@publicl.ptt.js.cn

魏之来　1980 年生,2002 年获安徽理工大学理学学士学位。现为硕士研究生。现主要研究领域为中文版面分析、模式识别、数字图像处理等。

杨静宇　1941 年生,教授、博士生导师。现主要研究领域为图像分析、图像理解、模式识别。