文章编号:1001 - 9081(2006)07 - 1587 - 03

基于空白条方向拟合的复杂文本图像倾斜检测

谢凤英[1],姜志国[1],汪 雷[2]
(1. 北京航空航天大学 图像中心,北京 100083; 2. 杜伊斯堡-埃森大学 国际工程学院,德国 杜伊斯堡 47057)
(xfy_73@buaa.edu.cn;xfy_73@sina.com)

摘 要:针对扫描背景不定且含有图表信息的复杂文本图像,提出了一种有效的倾斜检测方法。该方法首先通过对梯度图像的统计分析,自适应地选取到了包含文字的特征子区;在特征子区内,论文把文字行间的空白条带看作一条隐含的线,用优化理论计算出空白条带的倾斜角度,这也就是文本的倾斜角度。实验结果表明,该倾斜检测方法不受扫描背景、边界大小、文本布局及行间距等情况的限制,具有速度快、精度高、适应性强的特点。

关键词:复杂文本图像;倾斜检测;特征子区域;统计分析
中图分类号: TP391 **文献标识码**:A

Skew detection for complex document image based on blank bar directional fitting

XIE Feng-ying[1], JIANG Zhi-guo[1], WANG Lei[2]
(1. *Image Processing Center, Beijing University of Aeronautics and Astronautics, Beijing* 100083, *China*;
2. *International Studies in Engineering, University of Duisburg-Essen, Duisburg* 47057, *Germany*)

Abstract: An efficient skew detection algorithm was proposed for complex document image containing figure or form contents. The algorithm included three steps: Firstly, the text sub-region was selected adaptively according to the feature that the edges contained in text regions was stronger than those in non-text regions; Secondly, the blank bars between two text lines were extracted by blank blocks searching; Thirdly, the skew angle of blank bar was calculated by directional fitting, and this skew angle was just the document skew angle. The experiment results show that this algorithm is insensitive to the circumstances such as scan background, document layout and line spacing, and the skew angle can be detected rapidly and accurately.

Key words: complex document image; skew detection; feature sub-region; statistics analysis

0 引言

光学字符识别(Optical Character Recognition, OCR)是多年来一直活跃在图像处理与模式识别领域的一个重要研究方向,而文档图像的倾斜检测及校正则是 OCR 的预处理过程,角度检测的好坏直接影响了后续的字符识别率。目前,对文档倾斜检测的算法主要是根据文档所具有的不同特征提出来的,包括基于投影特征、基于线特征、基于频率特征、基于互相关特征等方法[1,2]。

基于投影特征的方法就是对文档图像进行不同角度的投影测试,在得到的一系列结果中提取最佳的投影结果,从而估算文档图像的倾斜角。这种方法比较适合于纯文本图像的倾斜检测,随着检测精度的提高,计算量也随之大大增加[3,4]。基于线特征的方法主要是 Hough 变换的方法,它将图像数据变换到参数空间,在参数空间完成线参数的计算,当然这个线与文档有相同的倾斜角,这种方法比较适合于含有线特征的图像的检测,而且随着检测精度的提高,这种方法的存储开销和时间开销都比较大[5,6];文献[2,7]中搜索图像中处在相同线上的一些关键点,通过线参数的拟合来检测倾斜角度,这种方法的关键是如何找到相同线上的点。基于频率特征的方法[8]是将图像经傅立叶变换到频域,频域空间密度最高的方向就是要求得的倾斜角度,这种方法的缺点也是计算的空间和时间的复杂度太高。基于互相关特征的方法[9]主要基于对每行文字固定行间距的研究,以得到最大化的倾斜角,这种方法同样计算复杂度高。

还有一些算法,如文献[10]中利用梯度矢量提取出线结构,然后通过角度直方图最高峰来计算文档的倾斜角。文献[11]中利用模糊游程的概念把图像分成文字区和图形区,并用最小二乘法计算出文档的倾斜角度。

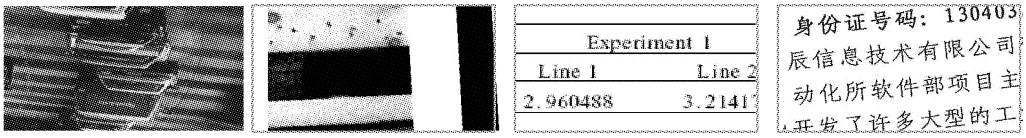
从以上算法可以看出,目前大多数算法都是根据图像中前景像素点(目标点)的纹理分布特征提出来的,而从文本图像的背景纹理分布出发研究问题的文献却很少。本文在总结上述算法的基础上,提出了一种根据文本行间背景纹理特征来检测倾斜角度的算法,通过拟合文字行间空白条的方向来达到对倾斜角的快速准确检测。同时针对含有图表信息、扫描背景不定的复杂文本图像,提出了一种特征子区自适应选取的预处理方法,大大提高了倾斜检测算法的速度和适应性,取得了良好的结果。

收稿日期:2006 - 01 - 24;**修订日期**:2006 - 03 - 21

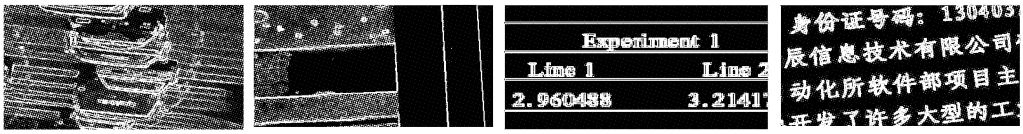
作者简介:谢凤英 (1973 -),女,黑龙江哈尔滨人,讲师,博士研究生,主要研究方向:医学图像处理、文档图像的分析和理解、遥感图像地面目标的检测; 姜志国 (1965 -),男,吉林梨树人,教授,博士,主要研究方向:医学图像处理、三维可视化、遥感图像处理; 汪雷(1982 -),男,湖北武汉人,硕士研究生,主要研究方向:文档图像的分析和理解.

1 子区选择

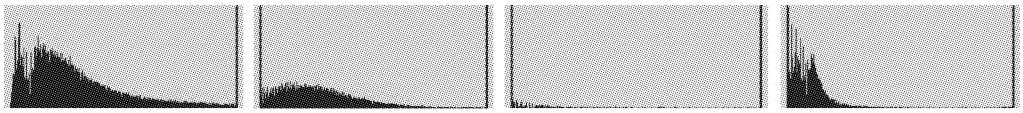
复杂文本图像包括以下情况:1)扫描图像的背景有可能是白边也可能是黑边;2)图像中含有图、表等内容;3)文档及边界大小情况不定,如图2中(a)和(c)所示。若图像的背景数据或文档中的图表信息与文字数据一起进行计算,则必然影响后面角度检测算法的准确性,同时较大的文档数据也会影响算法的速度。因此,提高倾斜检测的速度和准确度的最好办法就是选取特征子区,在子区中进行角度检测。目前,常用的子区选取方法是根据经验值在图像内部一定范围内随机选取满足一定条件的区域作为特征子区。这种方法对于黑白扫描边界类型未定且边界大小变化的情况很难达到自适应的效果,同时随机方法也不是全图最优的选取策略。



(a) 从左至右依次为1,2,3,4四个子区



(b) 对应(a)四个子区的梯度图



(c) 对应(b)四个梯度图的直方图(由于两端突起与其他位置对比悬殊,为了显示需要,此处使两端的最大值饱和了)

图1 文档中不同类型子区及其梯度图的统计分析

我们知道,图像的直方图反映图像灰度的统计特性,它表达了图像中取不同灰度值的面积或像素数在整幅图像中所占的比例,是图像中最基本的信息。因此可以利用直方图中所含有的信息寻找理想的区域作为特征子区。

图1是文档图像的不同类型子区及其梯度图上的统计分析,这些子区均来自实际扫描的图像,子区大小为372×232。根据图像梯度的性质,对应图像中灰度变化比较平坦的区域,其梯度图表现为暗区;对应图像中灰度变化比较剧烈的边缘地带,其梯度图表现为高亮带。因此,对梯度图进行直方图统计,则其直方图上会有两个突起(即直方图上非常陡的峰)分别位于低端和高端,这两个突起是直方图上的两个最大值,其值远远高于其他周围位置的值。表1是对这些直方图不同位置的高度值统计。

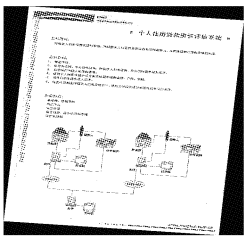
表1 图1中四个子区梯度图的直方图高度值(像素数)统计

位置(0~255)	子区1	子区2	子区3	子区4
低端突起	1194	48256	63973	39304
高端突起	7428	3562	10030	18496
其他位置	≤917	≤383	≤195	≤1114

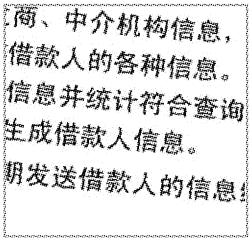
从表1中可以看出,文字区域(子区4)的高端突起(对应图像的边缘地带)的值远远高于图表子区(前三个子区)情况。这是因为:首先,文字与背景的对比度非常强,它的强边缘所占的比例要高于图形区域情况;其次,对于表格类型子区的情况,虽然它的边缘也很强,但其边缘所占的比例要低于文字区域。因此,对于一幅文档图像,如果一个子区在梯度图上的直方图高端突起值越高,那么这个子区是文字区域的可能性也就越大。

通过以上分析,我们把在梯度图上具有最大高端峰值特性作为最佳子区选择的依据。图2是两个最佳子区的选取实例。

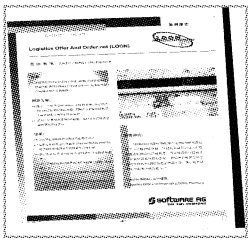
从本文子区选取的过程可以看到,这种子区选取方法不受扫描文档图像边界类型及大小的影响,不受文档中图表结构的影响,能够自适应地选取文字所在的区域,确保了倾斜检测算法的准确性。



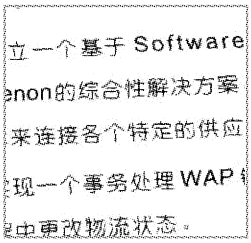
(a) 黑色前景扫描图像



(b) 从(a)中选取的最佳特征子区



(c) 白色背景扫描图像



(d) 从(c)中选取的最佳特征子区

图2 两个特征子区的选取实例

2 倾斜检测

空白条是文字行与行间的带状的空白区域,它与文档有相同的倾斜角度。因此,计算出特征子区中空白条的角度,也就计算出了文档的倾斜角度。

2.1 空白条带的搜索

先搜索行间的空白块,然后由这些空白块组成空白条带。设子区大小为 $W_r \times H_r$,子区内搜索空白条的大致步骤如下:

1)假定段宽为 L,将图像点阵从左至右按列向分成 $n = 1, 2, \cdots, N$ 个段,$W_r = L \times H$。

2)搜索 $n = 1$ 列所含有的空白块。搜索的方法为,用一个宽为 L 的线段从上到下搜索,当找到空白条时,将其记录下来作为空白块的上边沿;继续向下搜索,直到碰到一行有黑像素点的行时停止,并将其记录下来作为空白块的下边沿,至此找到该段内的一个空白块。然后继续进行下一个空白块的搜索,

直到找到该段所有空白块,把这些块记为 m_i, $i = 1, 2, \cdots, M$。

3) 搜索空白行。以前面每一个空白块 m_i 为起始,在其右侧上下交错位置搜索与其同行的空白块,将新搜索到的空白块作为新的起点,再搜索其下一个同行相邻的空白块,最后组成 m_i 所引领的空白条。

前面在搜索空白块时,段宽 L 的值是直接指定的。这个值选得太大会影响对大角度倾斜文档的检测,太小又会增加计算的复杂性,而且,段宽的选取也会影响图像倾斜检测的精度。为了达到对 L 的自适应选取,可以采取试探的方法来确定带宽 L,比如先用一个较大的值搜索,若每个空白行内包含的块数过少,则调整段宽 L 的大小,重新进行搜索。

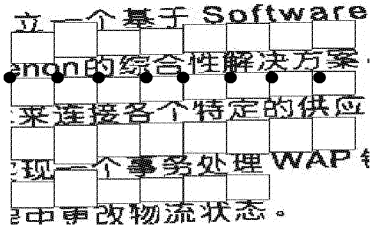


图 3　对图 2(d)的空白条搜索结果

2.2　倾斜角度的确定

从图 3 可以看出,一个空白条带由几个交错的块组成。这些块的左上角点就是空白条带上边沿上的点,如图 3 中的黑点,由这些点拟合直线,则拟合出的这条线与空白条有相同的方向,这也就是文档倾斜的方向。对某个空白条,假设它包含的空白块左上角的坐标为 (x_i, y_i), $i = 1, 2, \cdots, N$,并假设由这些点所拟合的最佳直线的方程为:

$$y = ax + b \tag{1}$$

根据一元线性回归方程,a 和 b 将满足:

$$\begin{cases} a = \dfrac{N\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \\ b = \dfrac{\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i x_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \end{cases} \tag{2}$$

则 a 的值就代表了文档的倾斜方向[12]。

由于一个子区内包含的空白条带可能是一行也可能是多行,具体计算时,可以选取最长的空白条带进行拟合,也可以对几个条带同时进行拟合并求平均,而且还应该排除那些过短或块不均匀的空白条参加角度计算。本文是选取具有均匀块且长度较大的几个空白条进行拟合求平均来计算角度的。

3　实验结果及分析

我们在 Pentium1.6,VC6.0 环境下上对投影法、Hough 变换法、本文方法三种方法进行对比分析。

实验 1　子区选择前后角度检测的时间分析

表 2　子区选择前后角度检测时间表(单位:s)

方法	投影法	Hough 法	本文方法
无子区选择	1.844	2.141	0.344
有子区选择	0.578	1.062	0.493

实验数据是一幅 大小 1 132 × 1 962 的文档图像,表 2 是子区选择有、无两种情况下三种方法角度检测的时间表(不考虑检测的正确与否)。

实验表明,子区选择降低了投影法和 Hough 变换方法的处理时间,但对本文方法却增加了处理时间,不过从整体上看,本文方法在时间上仍然优于其他两种方法。

实验 2　理想的复杂文本图像数据分析

理想图像是指用机器生成的图像并对其进行不等角度旋转而得的,图像中包含了图表等信息。参加测试的图像共 6 种、56 幅,倾斜角度小于等于 15°。表 3 是在子区选取的基础上对三种方法的统计结果,其中子区大小为 250 × 250。

表 3　三种方法对理想数据测试的统计表

方法	平均检测时间/s	平均绝对误差
投影法	0.119 393	0.002 925
Hough 法	0.071 179	0.006 593
本文方法	0.003 589	0.002 117

表 3 中,平均绝对误差为倾斜角的正切值。

实验表明,本文方法在处理时间和检测精度上都明显好于其他两种方法。

实验 3　实际扫描图像数据分析

实际扫描的复杂文本图像共 10 类、52 幅,包括不同背景扫描、边界及字符大小不等、含图表信息等情况,倾斜角度约在 10°以内。投影法、Hough 法和本文方法角度检测正确率分别为 88.5%, 71.2% 和 94.2%。正确率是指按照检测到的角度纠正倾斜文档后视觉上无倾斜的图像数与参加测试的图像总数之比。实验表明本文方法对实际扫描的复杂文本有比较强的适应性,通用性比较好。

4　结语

本文从文本行间背景纹理分布特性出发,根据文字行间空白条与文本倾斜角度的一致性,通过对空白条带方向的拟和实现文档倾斜角度的计算,算法实现简单,速度快。同时,针对复杂文本图像,为了降低数据复杂度,减少计算量,本文在梯度图上对包含文字、图、表及扫描背景等区域的灰度统计特征进行分析,实现了特征子区的自适应选取。实验结果证明,本文的整个倾斜检测过程不受扫描背景、边界大小以及图表等信息的限制,具有速度快、精度高、适应性强等特点。

另外,从实验 1 中可以看到,子区选取的过程虽然提高了投影法和 Hough 变换法的速度,但却降低了空白条带拟合法的速度。因此,如果能够在全图上准确提取出空白条带,将可以在确保正确率的前提下进一步提高倾斜检测的速度。

参考文献:

[1] 吴涛, 贺汉根. 一种快速的文本倾斜检测方法[J]. 计算机工程与应用, 2002, 38(5): 113 - 115.

[2] 张晓芸, 朱庆生, 曾令秋. 基于直线拟合的文本倾斜检测算法[J]. 计算机应用研究, 2005, 22(6): 251 - 253.

[3] LAM SM, ZANDY VC. Skew detection using directional profile analysis[A]. Proceedings of IAPR Workshop on Machine Vision and Apllication(MVA'94)[C]. Kawasaki, Japan, 1994. 95 - 98.

[4] 靳从, 魏之来, 杨静宇. 基于视窗的 OCR 页面图像倾斜检测方法[J]. 中国图象图形学报, 2004, 9(11): 1290 - 1293.

[5] 李政, 杨扬, 颉斌, 等. 一种基于 Hough 变换的文档图像倾斜纠正方法[J]. 计算机应用, 2005, 25(3): 583 - 585.

[6] LE DS, THOMA GR, WECHALER H. Automated Page Orientation and Skew Angle Detection for Binary Document Images[J]. Pattern Recognition, 1997, 27(10): 1325 - 1344.

(下转第 1597 页)

HBM_64 来表示。用 Synthetic 标准视频序列进行基于子块均值的块匹配,图 1 为视频序列中所有相邻图像帧多级块匹配的结果平均。可以看出,1-子块划分具有最低的 PSNR 值和最少的消耗时间,随着子块划分的增加 PSNR 值和消耗时间都随之增加。在多级块匹配的情况下,到达 4-子块划分的 PSNR 值与完全匹配误差函数的差值都大于 -0.5dB,4-子块划分后续的 PSNR 值都维持在与 4-子块划分相当的水平上。大量实验表明,在多级块匹配的情况下 4-子块划分符合最佳子块划分的要求,最佳子块划分所对应的消耗时间低于完全匹配误差函数。

2.2　多个匹配误差函数对比

用 26 个标准视频序列对基于不同匹配误差函数的多级(5 级)块匹配进行比较,分别为完全匹配误差函数、基于子块均值的匹配误差函数和基于子抽样的匹配误差函数[2]。对于每一个标准视频序列,对视频序列中所有相邻图像帧的块匹配结果进行平均,作为视频序列对应的运动估算参数。完全匹配误差函数用 FM(Full Matching)来表示;基于子块均值的匹配误差函数用 SM(Sub Block Mean)来表示,采用 4-子块划分(最佳子块划分);基于子抽样的匹配误差函数用 SS(Sub-Sampling)来表示,其每 4 个像素(X 和 Y 轴各 2 个像素)进行一次子抽样。图 2 给出了相应的峰值信噪比/视频序列曲线图和消耗时间/视频序列曲线图。

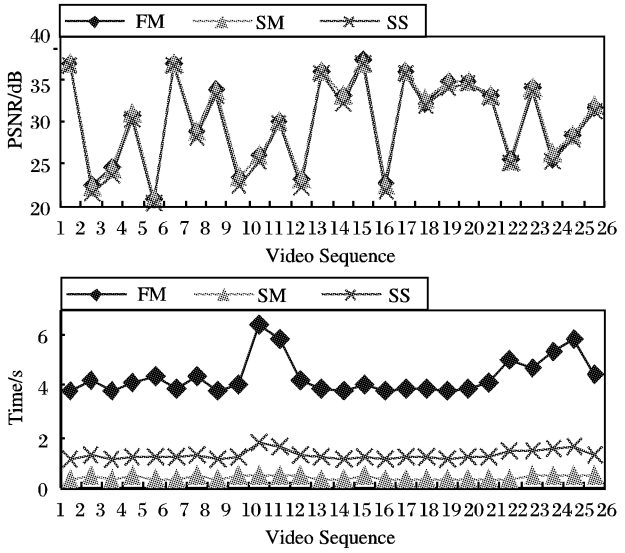


图 2　用 26 个标准视频序列比较不同匹配误差函数的多级块匹配

根据图 2 可以得出,基于子块均值的匹配误差函数对应的所有 PSNR 值与完全匹配误差函数相当地接近,符合最佳子块划分的 0.5dB 阈值要求,而基于子抽样的匹配误差函数对应的 PSNR 值与完全匹配误差函数的差距则大多超出了 0.5dB阈值。在消耗时间方面两种局部匹配误差函数都快于完全匹配误差函数,其中基于子块均值的匹配误差函数要成倍地快于基于子抽样的匹配误差函数。而且,基于子块均值的匹配误差函数相对其他两种匹配误差函数还有一个优点,就是对于不同的匹配图像其所消耗的时间大致相同,这非常适合于实时视频压缩等应用。

为了进一步地说明,我们对使用相同匹配误差函数的所有标准视频序列的运动估算参数进行平均,并制成表格(见表 1)。从表 1 可以得出,在 0.5dB 阈值的衡量下基于子块均值的匹配误差函数所对应的 PSNR 值与完全匹配误差函数相当,而基于子抽样的匹配误差函数所对应的 PSNR 值则不符合要求。基于子块均值的匹配误差函数所对应的消耗时间为完全误差函数的 9%,而相对基于子抽样的匹配误差函数则为 30%。实验表明,基于子块均值的匹配误差函数作为一种局部匹配误差函数,取得了与完全匹配误差函数相当的运动估算质量和快于完全匹配误差函数的运动估算速度;与基于子抽样的匹配误差函数相比具有更好的运动估算质量和更快的运动估算速度。

表 1　基于不同匹配误差函数的多级块匹配性能比较

匹配误差函数	PSNR/dB	Time/ms
FM	30.06	4407.27
SM	30.09	389.23
SS	29.53	1284.31

本文进一步的研究工作有两点:一是将基于子块均值的匹配误差函数与局部距离技术结合起来,以提供更快的匹配速度;二是将子块划分的恒常性与子抽样进行结合,就是对子块抽取单个像素代替均值参与匹配,并与基于子块均值的匹配误差函数进行比较。

参考文献:

[1] LIU B, ZACCARIN A. New fast algorithms for the estimation of block motion vectors[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1993, 3(2): 148 - 157.

[2] MOSEHETTI F, DEBES E. A fast block matching for SIMD processors using subsampling[A]. ISCAS 2000 - IEEE International Symposium on Circuits and Systems[C], 2000, IV-321 - IV-324.

[3] BEI C-D, GRAY RM. An improvement of the minimum distortion encoding algorithm for vector quantization[J]. IEEE Transactions on Communications, 1985, COM-33(10): 1132 - 1133.

[4] SRINIVASAN R, KAO KR. Predictive coding based on efficient motion estimation[J]. IEEE Transactions on Communications, 1985, COM-33(8): 888 - 896.

[5] NAM KW, KIM J-S, PARK R-H, *et al*. A fast hierarchical motion vector estimation algorithm using mean pyramid[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1995, 5(4): 344 - 351.

[6] VIOLA P, JONES, MJ. Robust real - time object detection[A]. IEEE ICCV Workshop on Statistical and Computational Theories of Vision[C]. Vancouver, Canada, 2001.

(上接第 1589 页)

[7] 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.

[8] PSTL W. Detection of Linear Oblique Structure and Skew Scan in Digitized Documents[A]. Proceeding of English International Conference on Pattern Recognition[C], 1986. 687 - 689.

[9] YAN H. Skew Correction of Document Images Using Inner Line Cross-correlation[J]. Computer Vision, Graphics and Image Processing: Graphical Models and Image Processing, 1993, 55(6): 538 - 543.

[10] KAO C-H, DON H-S. Skew Detection of Document Images Using Line Structural Information[A]. Proceedings of the Third International Conference on Information Technology and Applications[C], 2005.

[11] SHI Z, GOVINDARAJU V. Skew detection for complex document images using fuzzy runlength[A]. Proceedings of Seventh International Conference on Document Analysis and Recognition[C]. Edinburgh, Scotland, 2003. 715 - 719.

[12] 陈宝林. 最优化理论与算法[M]. 北京: 清华大学出版社, 1989.