基于投影的文档图像倾斜校正方法

张顺利[1,2],李卫斌[1],吉 军[2]
ZHANG Shun-li[1,2],LI Wei-bin[1],JI Jun[2]

1.咸阳师范学院 图形图像处理研究所,陕西 咸阳 712000
2.西北工业大学 现代设计与集成制造技术教育部重点实验室,西安 710072
1.Institute of Graphics and Image Processing,Xianyang Normal University,Xianyang,Shaanxi 712000,China
2.Key Lab of Contemporary Design & Integrated Manufacturing Technology of MOE,Northwestern Polytechnical University,Xi'an 710072,China
E-mail:slmmzhang@sina.com

ZHANG Shun-li,LI Wei-bin,JI Jun.Skew correction method for document image based on projection.Computer Engineering and Applications,2010,46(3):166-168.

Abstract: Aiming at the skew correction of document image,a novel skew correction method based on projection is proposed. Firstly,it projects the document image from different views using an efficient pixels traversal algorithm.Then,it calculates the sum of projection data and the skew angle can be determined by comparing these sums of different views.Since only very few part of the document image is projected during the procedure of projection,large amount of operations are saved.Based on the character of this method,a strategy of projection from rough to fine is proposed,which can greatly improve the speed and ensure the accuracy of detection.The experimental results show that the proposed method is very effective and can achieve very high accuracy.
Key words: image processing;skew correction;document image

摘 要:针对文档图像的倾斜校正问题,提出了一种新的基于投影的文档图像倾斜角检测方法。首先采用一种高效的像素遍历算法对文档图像从不同角度进行投影,然后对投影数据进行累加求和,通过比较不同角度下的累加和来确定倾斜角度。该方法在投影过程中只需对文档图像进行极少部分投影,因而大大减少了运算量。基于该方法的特点,提出了由"粗"到"精"的投影策略,在确保检测精度的同时大幅提高了检测速度。实验结果表明,方法非常有效,可以获得很高的检测精度。
关键词:图像处理;倾斜校正;文档图像
DOI:10.3778/j.issn.1002-8331.2010.03.050 **文章编号**:1002-8331(2010)03-0166-03 **文献标识码**:A **中图分类号**:TP391

1 引言

在将纸质文档资料通过图像采集设备进行光学扫描时,所得到的图像不可避免地存在一定程度的倾斜,这会给后续的文档图像的分析和处理带来困难。尤其是在对文档图像进行光学字符识别(Optical Character Recognition,OCR)时,由于文档图像的倾斜会降低文字的识别率,因此,有必要对文档图像进行倾斜校正处理。

倾斜校正的关键是如何高效、准确地检测出图像的倾斜角。目前,倾斜角检测的方法有多种,主要分为基于 Hough 变换的方法[1]、基于交叉相关性的方法[2]、基于 Fourier 变换的方法[3]、基于投影的方法[4-5]和 K-最近邻簇方法[6]等五类。其中基于投影的方法是最常用的倾斜校正方法,它是对文档图像进行不同角度的投影,得到若干投影图,再根据这些投影图的某些特征如均方差、第一特征矢量以及梯度等[7]的统计特性来求得文本倾斜角。但是由于传统投影方法需要对整个图像进行投影,且所需的投影方向多,因而计算量和复杂度都较高[8]。

针对传统投影方法的不足,考虑到文档图像的特点,提出了一种新的基于投影的快速、高精度的倾斜角检测方法,并通过实验对方法进行了验证。

2 算法原理

在文档图像中,文字行与行之间通常存在一定的间距。可以设想,当用一系列平行光线从不同角度投射到文档图像时,光线将被文字遮挡,其中只有与文字行相平行的光线才可以沿着行间隙最大量地穿过文档图像,此时的光线方向即是文档的倾斜角度。

基于上述思想,将文档图像的像素看作是边长为 δ 的正方形,光线看作是由一系列宽度为 τ 的光束组成,通常取 $\tau=\delta$。为了方便计算,令文字、表格等所覆盖部分的像素值为 1,未被覆盖部分的像素值为 0。实际中所获得的文档图像通常为灰度图

基金项目:陕西省自然科学基金(the Natural Science Foundation of Shaanxi Province of China under Grant No.2007D22,2009JQ8017);陕西省教育厅专项基金(No.09JK810)。
作者简介:张顺利(1973-),男,博士研究生,副教授,主要研究方向:计算机图形图像处理;李卫斌(1976-),男,博士后,教授,主要研究方向:计算机图形图像处理;吉军(1982-),男,博士研究生,主要研究方向:计算机辅助设计与制造。
收稿日期:2008-07-19 **修回日期**:2008-09-01

像,为了去除噪声和简化计算,需要对文档图像进行二值化预处理。作如下的规定:如果一个像素的中心位于光束内,则光束经过该像素;否则,不经过该像素。这样,当一条光束投影到文档图像时,一旦经过值为 1 的像素,则这条光束将被遮挡,记投影值为 1;如果光束所经过的所有像素值均为 0,则光束完全穿过文档图像,记投影值为 0。将同一角度下的投影值进行累加求和,根据上述规定,如果光束越接近文档图像的倾斜角,则累加和越小,反之则越大。因此可以对文档图像进行不同角度的投影,并对投影值进行累加求和,根据和的大小最终确定文档图像的倾斜角。

3　算法实现

设文档图像由 $W\times H$ 个边长为 δ 的像素组成,如图 1 所示。图像宽为 $WIDTH$,高为 $HEIGHT$,左下角位于坐标原点 O。对所有像素按从左到右、从上到下的顺序进行编号,依次为 0、1、…、$HW-1$,对应的图像灰度值为 $f[i]$,其中 $0\leqslant i\leqslant HW-1$。由上述分析可知,该文方法的关键是实现光束与像素的快速遍历。在文献[9]中,提出了一种射束与像素的遍历算法,该算法通过增量计算,而且主要涉及到加、减法运算,因而具有很高的效率。

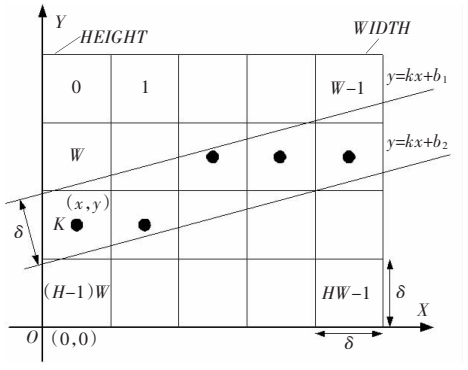


图 1　光束投影示意图

考虑到多数情况下,文档图像的倾斜角度不会过大。不失一般性,假定光束的斜率 k 满足 $0\leqslant k<1$,其上边界直线方程为 $y=kx+b_1$,下边界直线方程为 $y=kx+b_2$。由几何关系不难得到如下性质:

性质 1　若一个像素在光束内,且其正上方像素也在光束内,则其右上方像素必然也在射束内。

性质 2　若一个像素在光束内,且其正右方像素也在光束内,则其正上方像素必然不在射束内。

性质 3　若一个像素在光束内,且其正右方和正上方像素均不在光束内,则其右上方像素必然在光束内。

当 $0\leqslant k<1$ 时,由于光束在 X 方向比 Y 方向变化要块,所以沿 X 方向步进。在步进过程中,当确定一个像素 K 在光束内时,根据上述性质,下一个要遍历的像素按如下规则来确定:

(1)首先判断正右方像素 $K+1$ 是否在光束内,若在,则遍历该像素。

(2)若正右方像素 $K+1$ 不在光束内,则判断正上方像素 $K-W$ 是否在光束内,若在,则依次遍历正上方像素 $K-W$ 和右上方像素 $K-W+1$。

(3)若正右方像素 $K+1$ 和正上方像素 $K-W$ 都不在光束内,则遍历右上方像素 $K-W+1$。

在遍历像素之前,需要计算文档图像中光束经过的初始像素 K 及其中心坐标 (x,y)。为了计算每个投影角下的投影,定义一个数组 $proj[N]$,其中 N 为每一投影角下的光束数,通常取 $N=H$,并令 $\delta=1$。下面给出某一角度下的投影运算的伪代码:

```
for(i=0;i<N;i++)
{ 计算第 i 条射束对应的 b1、b2、K 及其中心坐标(x,y),F1=kx+b1;
F2=kx+b2;proj[i]=0;
  if(f[K]>0){proj[i]=1;continue;}
  do{
    if(y>=F2+k)//判断正右方像素是否在光束内
    {
      if(x<WIDTH-1)
      { K++;
        if(f[K]>0){proj[i]=1;break;}
          x++;F1=F1+k;F2=F2+k;
      } else break;
    }
  else if(y+1<=F1)//判断正上方像素是否在光束内
    { if(y<HEIGHT-1)
      { K=K-W;
        if(f[K]>0){proj[i]=1;break;}
        y++;
      }else break;
      if(x<WIDTH-1)
      {K++;
        if(f[K]>0){proj[i]=1;break;}
        x++;F1=F1+k;F2=F2+k;
      }else break;
    }
  else //右上方像素在光束内
    {
        if(x<WIDTH-1 && y<HEIGHT-1)
        { K=K-W+1;
          if(f[K]>0){proj[i]=1;break;}
          x++;y++;F1=F1+k;F2=F2+k;
        }else break;
    }
  }while(x<WIDTH && y<HEIGHT);
}
```

上述算法的特点在于投影过程中,只要遇到某个像素的值非 0,即停止遍历,并令投影值为 1。和传统方法相比,可以大大节省运算量;另外,该算法为增量运算,所以效率很高。在得到某一角度下的投影数据后,对投影数据 $proj[N]$ 进行累加求和,并将累加和保存在数组 $sum[M]$ 中,M 为投影角总数。用同样的方法,改变投影角的间隔 δ 再计算。最终数组 $sum[M]$ 中保存了所有投影角下的投影累加和。

根据上述讨论,文档图像的倾斜角应是数组 $sum[M]$ 中取得最小值时对应的角度。但是,如果投影角的间隔 δ 取得很小,可能会出现多个连续角度下的投影累加和同时取得最小(比如从 n_1 到 n_2),这时可以取倾斜角度为 $(n_1+n_2)/2$。

对于 $-1<k<0$ 时的情形,可以根据上述讨论由对称性得到。

4 实验结果及分析

选取一幅标准的没有倾斜的文档图像,并人为旋转不同的角度得到倾斜图像,通过这些倾斜的文档图像来测试该文算法的检测精度。将 Hough 算法、交叉相关算法及该文算法分别应用于不同倾斜角的同一文档来进行对比。测试计算机配置为 Pentium 4 2.8 GHz CPU、512 MB DDR 内存。实验过程中投影角的间隔 δ 取 0.1°。表 1 给出了三种算法检测结果的对比。

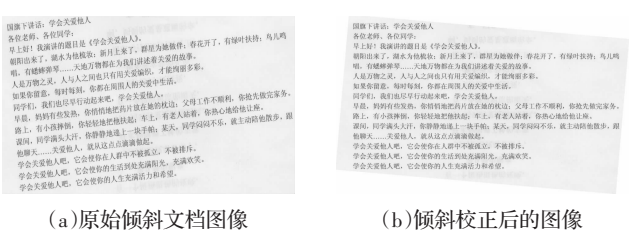
表 1 三种算法的检测精度对比 (°)

倾斜角	Hough 算法	交叉相关算法	该文算法
−3.8	−4.0	−3.82	−3.75
−2.8	−2.9	−2.89	−2.8
−1.8	−1.9	−1.83	−1.8
−0.8	−1.1	−0.86	−0.8
0.1	0.0	0.10	0.05
0.2	0.1	0.10	0.2
1.2	1.0	1.07	1.2
2.2	2.1	2.13	2.2
3.2	3.0	3.09	3.2
4.2	4.1	4.13	4.2
最大绝对误差	0.3	0.13	0.05
最小绝对误差	0.1	0.0	0.0
平均误差值	0.15	0.068	0.01

由表 1 可以看出,该文算法的最大绝对误差为 0.05°,最小绝对误差达到 0°。如果 δ 进一步取 0.01°,则最大绝对误差将不超过 0.005°。平均误差反映了算法精确度,其值越小表示精确度越高。Hough 算法的平均误差为 0.15°,交叉相关算法为0.068°,而该文算法仅为 0.01°。实验证明,该文算法具有很高的精度。

由于该文方法是基于投影的,因此检测精度很大程度上取决于投影角的间隔 δ。若 δ 过大,则检测速度快,但精度低;反之,则速度慢但精度高。为了同时提高倾斜角检测的速度和精度,提出了一种由“粗”到“精”投影策略,即先用大角度间隔进行投影(通常取 1°),以确定大致的倾斜角度 n;然后在[n−0.5,n+0.5]角度范围内以小角度间隔进行投影(如 0.1°或 0.01°),这样可以大幅减少运算量,同时确保足够的精度。

以一幅实际扫描的 920×508 文档图像为例,如图 2(a)所示,先用该文算法在[−8,8]角度范围内以间隔 0.01°进行投影,检测到倾斜角为 4.975°,耗时 2.235 s。然后采用由“粗”到“精”的投影方法进行测试,其中大角度间隔为 1°,检测到的倾斜角为 5°,接着用小角度间隔 0.01°在[4.5,5.5]角度范围内投影,计算出倾斜角为 4.975°,实际耗时 0.628 s。可见,改进后的方法在保持检测精度的同时将检测速度提高了 2.5 倍多。图 2(b)为倾斜校正后的文档图像。



(a)原始倾斜文档图像 (b)倾斜校正后的图像

图 2 倾斜校正实例

为了进一步测试该文算法的效率,选取 20 副 A4 页面大小的文档图像,倾斜角度从−5°~5°不等,以 300 dpi 的分辨率进行扫描输入。分别使用 Hough 算法、交叉相关算法及该文算法进行倾斜角的检测,其中该文算法中的大角度间隔为 1°,小角度间隔取 0.1°。对比实验结果如表 2 所示。

表 2 三种算法的检测时间对比 s

	Hough 算法	交叉相关算法	该文算法
平均处理时间	3.164	2.621	2.275

由表 2 可以看出,该文算法倾斜角检测的平均处理时间为 2.275 s,而 Hough 算法的平均处理时间为 3.164 s,交叉相关算法的平均处理时间为 2.621 s。可见,该文算法具有较快的处理速度。

5 结论

文章提出了一种新的基于投影的文档图像倾斜校正方法。该方法采用一种高效的像素遍历算法对文档图像进行部分投影运算,克服了传统投影方法计算量大的不足。基于该文方法的特点,提出了一种由“粗”到“精”的投影方法,在保证精度的同时大大提高了倾斜角的检测速度。实验结果表明,该文方法可以达到很高的精度,这为文档图像的后续处理奠定了良好的基础。

参考文献:

[1] 周冠玮,平西建,程娟.基于改进 Hough 变换的文本图像倾斜校正方法[J].计算机应用,2007,27(7):52-57.

[2] Gatos B,Papermarkos N,Chamzas C.Skew detection and text line position determination in digitized documents[J].Pattern Recognition,1997,30(9):1505-1519.

[3] Pstl W.Detection of linear oblique structure and skew scan in digitized documents[C]//Proceedings of the 8th International Conference on Pattern Recognition,Paris,France,1986:487-489.

[4] Ciardiello G,Scafur G,Degrandi M,et al.An experimental system for office document handling and text recognition[C]//Proceedings of Ninth International Conference on Pattern Recognition,1998:739-743.

[5] Baird H S.The Skew angle of printed documents[C]//SPSE 40th Annual Conference and Symposium on Hybrid Imaging System,1987:739-743.

[6] 何希平,李云峰,朱庆生.彩色文档图像的倾斜自动校正算法[J].中国图象图形学报,2006,11(3):367-370.

[7] 卜飞宇,刘长松,丁晓青.灰度名片图像快速倾斜检测和校正方法[J].中文信息学报,2004,18(1):62-69.

[8] 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.

[9] 张顺利,张定华,王 凯,等.一种基于 ART 算法的快速图像重建技术[J].核电子学与探测技术,2007,27(3):479-483.