IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. VII (Sep – Oct. 2014), PP 136-139
www.iosrjournals.org

Skew Detection based on Bounding Edge Approximation

P. Malathi[1]
[1](*Master of Computer Applications Department, Dr. Ambedkar Institute of Technology, Visvesaraya Technological University, India)*

Abstract: *Any paper document when converted to electronic form through standard digitizing devices, like scanners, is subject to a small tilt or skew. With recent advances of hand-held devices such as cell-phones, Personal Digital Assistants (PDA), etc. having built-in digital cameras, a new trend of document capturing has emerged. Because of the non-contact nature of digital cameras attached to these handheld devices, acquired images very often suffer from skew distortion. A de-skewed document allows a more compact representation of its components, particularly text objects, such as words, lines, and paragraphs. This simplified representation leads to more efficient, robust, as well as simpler algorithms for document image analysis including optical character recognition (OCR). This paper presents a new method for automatic skew detection in images using heuristic approach .The proposed algorithm is fast as well as independent of the scripting language.*
Keywords: *Skew,Grayscale, Deskew,Rotation,Translation*

I. Introduction

The Document processing has gained the attention of researchers and industries in recent times due to its immense potential in commercial applications. The input document images can be obtained from a large variety of media, such as journals, newspapers, magazines, memos, etc. The format of document image can be digitally created, faxed, scanned, machine printed, or handwritten, etc.

The main stages in document processing are-i) Optical Scanning ii)Preprocessing iii) Document Analysis as shown in Fig 1.

i) Optical scanning is a process through which a digital image is captured from the original document.
ii) The preprocessing stage which includes thresholding, binarizing, filtering, skew detection, gap filling, segmentation and so on can make the initial image more suitable for later computation.
iii) Document Analysis is to recognize the text and graphics components in images, and to extract the intended information as a human would. Two categories of document image analysis can be defined . Textual processing deals with the text components of a document image. Some tasks here are: recognizing the text by optical character recognition (OCR), determining the skew (any tilt at which the document may have been scanned into the computer), finding columns, paragraphs, text lines, and words. Graphics processing deals with the non-textual line and symbol components that make up line diagrams, delimiting straight lines between text sections, company logos etc.

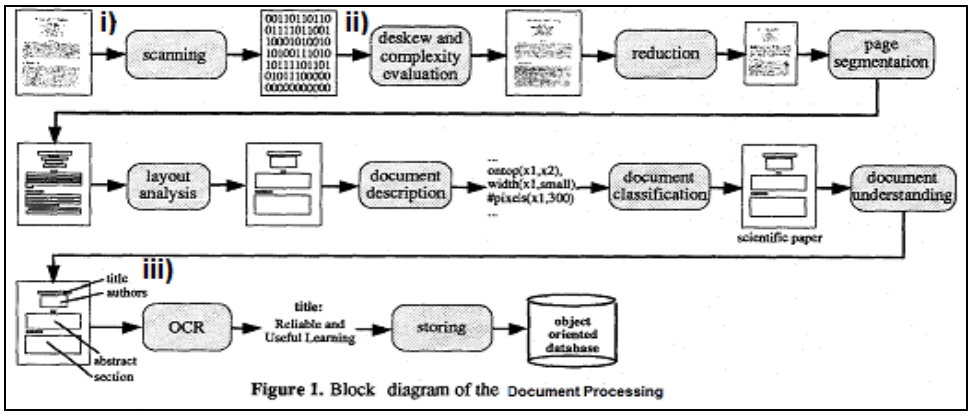

Figure 1. Block diagram of the Document Processing

II. Related Work

The skew of the scanned document image specifies the deviation of the text lines from the horizontal or vertical axis. One example of a process which is spoilt by skew is the use of horizontal and vertical projection profiles. Projection profiles have many applications in document image processing and they rely on horizontal and vertical lines being aligned to the axes. A number of good skew estimation algorithms are available in the literature. However, the time required to estimate the skew angle is still an important issue. These algorithms

focus on skew estimation at the layout analysis stage or the text processing stage . Most of these algorithms may be classified into four groups:- i) Hough transform [1,2,3,4,5,6] ii) Projection Profile[7,8] iii) Clustering of nearest neighbor[9,10,11,4] iv) Fourier transformation [12]

i) In Hough transform, the points in the Cartesian coordinate system are described as a summation of sinusoidal distribution: $p = x\cos\theta + y\sin\theta$, the skew angle is calculated on the basis that at the skew angle the density of Transform spaces is maximum. The major problem with Hough transform is the massive computational cost. Modifications of the Hough transform approach are also proposed that discard irrelevant pixels in order to obtain accurate transform peaks and reduce the processing time. However, the algorithm is not that fast compared to the proposed one.

ii) Projection profile is generally a histogram of the number of dark pixels in horizontal scan lines of a text. The document is projected at different angles. Then peaks (due to text line positions) and troughs (due to inter-line gap positions) are identified. The angle which gives the maximum difference between the peaks and troughs is accepted as the skew angle. The method, being computationally expensive, has been improved by proposing a quick convergence of this iterative approach. Projection profile methods are, in general, well suited to estimate skew angle within ±10 degrees.

iii) Clustering of nearest neighbors has been used by Hashizume [9] and Jiang et al. [10]. Hashizume computed the directions of the nearest neighbors of all the pixels in each connected component of the image. A histogram of the directions is constructed which indicates the skew angle from its peak. This method is generalized by O'Gorman [11]. Clustering of pixels along the mean line and base line has been successfully used by Pal and Chaudhuri[4] where component labeling is used to compute the average height of the characters.

iv) Fourier Transformation [12] works on the basic principle that skew angle is the one at which concentration of spectrum is biggest for the document.

III. Proposed Method

While the proposed skew estimation algorithm focus is at the preprocessing stage. Skew angle is determined based on well-defined structures of the text portion in a document. Our algorithm, like others, expects documents mostly filled with text lines. However, it can handle documents with a moderate amount of graphics. The algorithm has the additional advantage of being script independent.

The present algorithm is based on the observation that good skew features can be extracted from the borders of the image entities. For most characters, especially Latin ones, bottom borderlines can provide excellent measurement of skew angles. The present skew detection algorithm identifies base line of text at the border and extracts skew angle information based on it. The overall principle of this approach is more of a heuristic method. It is used for deskewing textual images at the page level. It is preferred to apply this algorithm on grayscale image after binarization and after border noise removal stage. Even if border noise removal facility is not used, it still can be applied if atleast one side of the image is without border noise problem. It is independent of the scripting language, the font size and resolution used in the textual image. It is applicable for both handwritten and printed documents. It works on pages with header, footer, page numbers and also otherwise. It will be the best and the efficient algorithm if the document has a header or the footer line.
Skew Detection based on bounding edge approximation algorithm is as follows. Here we find the bounding edge at the bottom of the document.

3.1 Algorithm

1. Initialize the probable set of pixels to form the bounding edge P to null.
2. Skip all the rows where there is no foreground pixels.
3. Check for the feasible bounding segment regions as follows-

```
For each row i with a foreground pixel k
    For each column j across the row i
        If P = Ø
        Then P = {k}, start a new segment
        If count(P) = width m of the image
        Then stop
        If k is just above or just below k-1
        Then P = P + {k}
        Else reject it and start a new segment
        If k is at the beginning of a new segment
        Then P = P + {k}
```

4. Find the threshold of all the segments formed in P using a heuristic method.
5. Remove all the pixels from P which are above or below this threshold to form base line set B.

6. Apply least square method to obtain the best line fit on B to get the bounding edge E. It also gives the slope of the edge E.
7. Slope of the edge E approximates the skew angle Θ of the image.
8. Rotate the image about its centre by angle Θ. Θ can be positive or negative to produce clockwise or anticlockwise rotation.

The above algorithm was implemented on Microsoft C#. Net. A sample of the results is shown in the figure below.

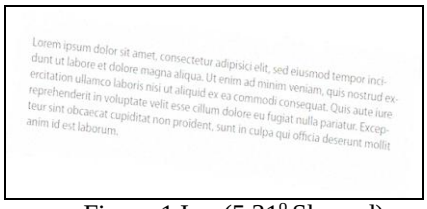

Figure 1 Img(5.31^o Skewed)

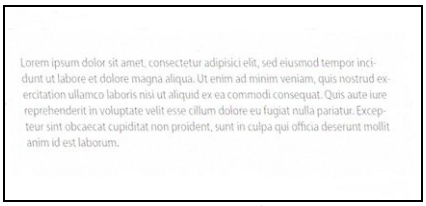

Figure 2 Img(5.3071^o Deskewed)

If the image suffers from multiple skew (i,e the contents are not perpendicular along the two axis) then after deskewing, the image looks sheared along the horizontal axis. There is no standard formula for unshearing transformation. So this can be followed by unshearing algorithm as follows-

1. Use the above algorithm modified to find the vertical edge V. It gives the vertical slope V_m and intercept b.
2. For each row across the height of the image
   Find the translation factor tx along the horizontal axis using V_m and b.
   Apply tx for each pixel along the width of each row.

   (However it does introduce some staircase effect on the contents of the image)

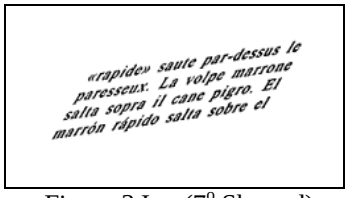

Figure 3 Img(7^o Skewed)

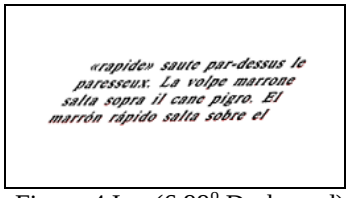

Figure 4 Img(6.99^o Deskewed)

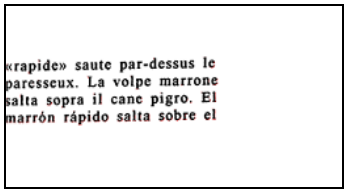

Figure 5 Img(Translated)

3.2 Formulation of Best Fit Line using Least Square method

$$m = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Slope m is given by where *n* is the total number of data points.

Compute the *y*-intercept of the line by using the formula: $b = \bar{y} - m\bar{x}$

Where $\bar{y}$ and $\bar{x}$ are the mean of the *x*- and *y*-coordinates of the data points respectively.

3.3 Complexity

If the document is a matrix of pixels of the order n*m then the algorithms time efficiency is of the order |n+m| for detecting the skew. For unskewing it is of the order |n*m|. If transformation has to be applied then it is of the order |n*m|

IV. Implemetation Results

DISEC 2013 was the first Document Image Skew Estimation Contest held within ICDAR 2013. [13] The benchmark dataset available on the DISEC homepage consists of binarized images. Of these two test sets are used here for testing the algorthm.The first dataset consists of the grayvalue images of the DISEC benchmarking dataset. The second dataset consists of a synthetic dataset which has been reproduced from the Paper of Boris Epshtein: Determining Document Skew Using Inter-Line Spaces, ICDAR 2011. The dataset is produced from 8 images, to which blur is added and Gaussian noise is applied. The angle ranges from 0 to pi radians.For each dataset each image is rotated 10 times within an angle interval of -15 to 15 degree. The GT

angle is encoded in the file name (comparable to the DISEC contest) . Table-I shows the summary of the outcome.

Table- I

Dataset	Samples taken	Accuracy range	Average error %
DISEC 2013 grayvalue images dataset	50	0.001º to 3.3821º	1.32
Synthetic dataset	80	0.001º to 2.166º	0.8

V. Conclusion

Thus deskewing is one of the important preprocessing stage in the document analysis. As the principle applied can greatly influence and affect the efficiency of all the other stages, accuracy of deskewing is of great importance. For now, the algorithm achieves reasonable accuracy for the skew angle, depending on the quality of the input data. The proposed algorithm works well for small skew angles (less than 10 degrees). Although skew detection methods without such restrictions are reported in the literature [14,4],in reality the skew seldom exceeds ±5 degrees[15]. Further work will be done in order to obtain better measure of the skew detection success. In future the algorithm will be applied for a variety of standard case study datasets like UAW-III dataset, PRIMA dataset and precise comparison statistics with the other existing standard algorithms will be worked out. The algorithm can be improved and adapted to apply at the line level deskewing of textual images of the handwritten documents also.

References

[1] Richard O. Duda,Peter E. Hart, April 1971. "Use of the Hough Transform to deterct lines and curves in pictures". Technical Note 36, AI Center.

[2] Srihari, S.N. and V. Govindaraju, 1989. "Analysis of textual images using the Hough transforms". Machine Vision Applications, 2: 141-153. DOI: 10.1007/BF01212455.

[3] Le, D.S., G.R. Thoma and H. Wechsler, 1994. Automaticpage orientation and skew angle detection for binary document images. Pattern Recognition, 27: 1325-1344.

[4] Pal, U. and B.B. Chaudhuri, 1996. An improved document skew angle estimation technique. Pattern Recognition Lett., 17: 899-904. DOI: 10.1016/0167-8655(96)00042-6

[5] Yu, B. and A.K. Jain, 1996. A robust and fast skew detection algorithm for generic documents. Patt. Recog.,29: 1599-1629. DOI: 10.1016/0031-3203(96)00020-9

[6] Tian Jipeng, G.Hemantha Kumar, H.K. Chethan : "Skew correction for Chinese character using Hough transform". International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Image Processing and Analysis.

[7] Hou, H.S., 1983. Digital Document Processing. Wisely NewYork, ISBN: 0471862479.

[8] Akiyama, T. and N. Hagita, 1990. Automated entry system for printed documents. Pattern Recognition, 23: 1141-1158. DOI: 10.1016/0031-3203(90)90112-X

[9] A. Hashizume, P.S. Yeh, A. Rosenfeld: A method for detectingthe orientation of aligned components. Pattern Recognition Lett., 4:125–132, 1986

[10] X. Jiang, H. Bunke, D. Widmer-Kljajo: Skew detection of document images by focused nearest-neighbor clustering. In: Proc. ICDAR99 5th Int. Conf. on Document Analysis and Recognition, pp. 629–632, Bangalore, September 20– 22, 1999

[11] O'Gorman, L., 1993. The document spectrum for page layout analysis. IEEE Trans. Patt. Anal. Mach. Intell., 11:1162-1173. DOI: 10.1109/34.244677.

[12] Omar, K., A. Ramli, R. Mahmod and M. Sulaiman, 2002. Skew detection and correction of jawi images using gradient direction. Journal of Tech., 37: 117-126.

[13] Robert Sablatnig, CVL, Vienna University of Technology: www.caa.tuwien.ac.at/cvl/research/skew-database/

[14] O. Okun, M. Pietikainen, J. Sauvola: Document skew estimation without angle range restriction. Int. J. Doc. Anal. Recognition 2:132–144, 1999

[15] I.T. Phillips, S. Chen, R.M. Haralick: Cd-rom document database standard. In: ICDAR93, pp. 478–483, 1993