Skew Detection and Skew Correction in scanned Document Image using Principal Component Analysis

Basavanna M, S. S. Gornale

Abstract—Skew detection and correction in a document image processing has become an imperative technology in the field of Digital Image Processing and pattern recognition which will take vital role in automation of office documentation. Skew is in exorable introduced into the scanned document during scanning, and it has direct effect on the reliability and efficiency of the segmentation and feature extraction stages for various applications. Hence, skew detection and correction in document images are critical steps before layout analysis. In this work a novel skew detection method is presented for skew detection and correction scanned document Images using Principal Component Analysis (PCA). The experiments which we have carried out on scanned document images are satisfactory and the performance of the proposed method is compared and analyzed in detail and the promising results and findings are presented.

Index Terms— Skew detection, skew correction, Principal Component analysis, Document image processing, PCA, Multi-lingual scanned document images, document image.

1 Introduction

A Document image contains different structural units such as text, graphics, tables, images, etc. Documented image analysis generally requires document image which does not include any skew. The skew of the document image can be global (all document's blocks have same orientation), multiple (document's blocks have different orientation) or non-uniform (multiple orientation in a text line) and also includes accurate measure of skew, within-line and between–line spacing's. Skew is an inevitable process and its detection is an important issue for document recognition systems because it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages. It includes the information about the technique used to detect skew which are introduced during the scanning of the documents.[11],[12],[13]. The skew of scanned document image specifies the deviation of the text lines from the horizontal or vertical axis. As a result, a document skew detection and correction is required before the actual analysis of the document starts.[1][2].

There are various methods that are existing for skew detection and skew correction based on the feature extractions methods for a scanned document images and they are divided into Irrelevant features Weakly relevant features, Strongly relevant features and dimension reduction methods which are divided again into Feature transformation, Feature extraction and the optimal subset includes all of the strongly relevant and weakly relevant but non-redundant features.[3],[13],[14],[15]. This work aims to remove the skew efficiently in a document image and the results are compared with the existing methods and found to be satisfactory and the performance of the proposed method is compared and analyzed in detail and the promising results and findings are presented.

The organization of the paper is as follows. We present the related work and their algorithm in section 2. The proposed methodology is presented in section 3. Experimental results are reported in section 4. Discussion based on experimental results is given in section 5. Conclusion and future work at the end followed by references.

2 Related Work

Sepideh Barekat Rezaei et. al., have proposed method for calculation of skew in a document image and which is based on connected components by smoothing algorithm and calculates the document skew by finding the orientation of the minimum area bounding rectangle of one or several connected components. The algorithm works in detecting document skew for a variety of documents with different levels of complexity [3].

Joost Van Beusekom, et.al. (2009) have proposed a one step method of skew and orientation detection using a line finding algorithm and which is resolution independent [4].

Manjunath Aradhya et. al. (2007) have proposed a novel skew detection method for binary document images by selected characters of the text which may be subjected to thinning and Hough transform to estimate skew angle accurately. This method also works various types of documents such as documents containing English Documents, Journals, Text-Book, Different Languages and Document with different fonts, documents with different resolutions [5].

- *Basavanna M, Post Graduate Department of Computer Sciecne, Govt. College (Autonomous), Mandya-Karanataka, India Mobile No. 9448536782. E-mail: basavanna_m@yahoo.com.*
- *Shivanand S Gornale, Department of Computer Science, School of Mathmetics and Computing Science,Rani Channamma University, Belagavi-Karanatak-India, E-mail: shivanand_gornale@yahoo.com. Mobile No. 9739364083 .*

S.Lu et. al. (2006) have worked on language and orientation detection using the distributions of the number and position of white to black transitions of the components in the line. The performance of their method is reported on a partially non-public dataset and achieves a success rate of 98.2% for documents with at least 12 text lines [6].

K Omar et. al. (2009) have proposed skew detection and correction technique for Arabic document images based on centre of gravity; it involved inscribing the text in the document by an arbitrary polygon and derivation of the baseline from polygon's centroid [7]

Zhixin Shi et al. (2003), have worked on skew angle estimation and correction of a document page is an important task for document analysis and optical character recognition (OCR) applications. And they have proposed a novel method for complex document skew angle estimation. The method is able to detect the per-dominant direction of the text in the document image. The text can be handwritten, machine printed or mixed [8].

P Shivakumara et al, (2005) have proposed a novel technique for estimation of skew in binary text document images based on linear regression analysis The method uses the boundary growing approach to extract the lowermost and uppermost coordinates of pixels of characters of text lines present in the document, which can be subjected to linear regression analysis (LRA) to determine the skew angle of a skewed document. [9]

A. Alaei et. al. (2011) have worked on automated page orientation and Skew Angle Detection for Binary Document Images A new and fast approach is advanced herein whereby skew angle detection takes advantage of information found using the page orientation algorithm. Page orientation is accomplished using local analysis, while skew angle detection is implemented based on the processing of pixels of the last black runs of binary image objects. The detection accuracy can be improved by minimizing the effects of non-textual data. The performance of the Hough transform used to detect the skew angle is sped up using data reduction procedure and page orientation information.[10]

3 Methodology

Many methods are proposed and analyzed for the skew detection and correction which are based on Principle Component Analysis (PCA), Project based, Hough-transform-based, Nearest –neighbor-based method and Cross-correlations-based methods by limiting their applications angle of skew in the images and still there is room for developing a robust algorithm for skewed images with texts, graphics, scenes and many more in the document images [16],[17]. The steps involved in the skew detection and correction is shown in below figure-1. In proposed methodology, we are inputting the image which contains skew, using MATLAB we have converted the original image into binary image, after generating the binary image we are finding the threshold value by plotting the histogram. Threshold value is useful for separating the fore ground and back ground of the given image. After that we are finding the sobel edges to the binary image. Then we are applying Principal Component Analysis (PCA) to the first axis (either x-axis or y-axis). PCA shows the skew angle of the image; means skew is detected, after skew is detected, next step is correction of the skew, After skew is detected we are adding or subtracting the angle(skew) to the original image and rotating the image to the requirement. This image is our output image where skew is eliminated.

3.1 Principal Component Analysis

The Principal Component Analysis (PCA) is a standard tool in modern data analysis - in diverse fields from neuroscience to computer graphics - because it is a simple, non-parametric method for extracting relevant information from confusing data sets. It is also an efficient tool in identifying and highlights the similarities and/or differences in the data [18], [19].

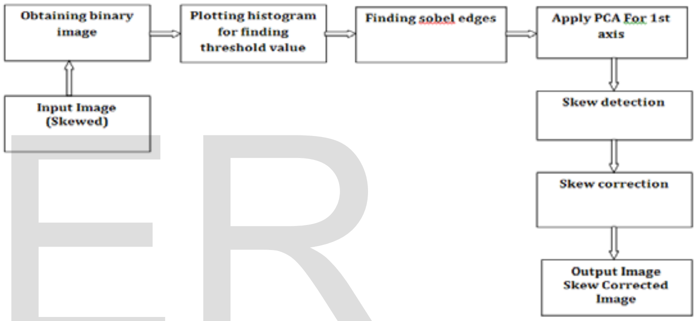

Figure-1: Skew detection and correction

4. Experimental Analysis and Discussion

In this paper we have carried out experiments on ICDAR-2003 database and also on our own database which is created using mobile camera with 2 mega pixel resolution in which image contains the skew. The skew detection and skew correction is done through the following algorithm-1.

```
Algorithm-1
Input image: Skewed image.
Output image: Skew corrected image.
Step-1: Input the skewed image.
Step-2: Convert the gray image to binary image.
Step-3: Plot the histogram for the finding the optimal threshold value.
Step-4: Find the edges using soble edge method.
Step-5: Apply the Principal Component Analysis (PCA) for first axis.
Step-6: Detect the skew.
Step-7: Display the corrected the skew.
```

Firstly, the experiment is carried out on standard skewd images database available on internet using the above algorithm on 150 sample images, as a sample of a result the below figure-2 show the various steps involved in skew detection and skew correction. In this sample of image the 38.07° angle is detected and the remaining images with skew corrected and displayed in different angles.

Secondly, the experiment is carried out on our own database using the above algorithm on 155 sample images, as a sample of a result, the below figure-3 shows the various steps involved in skew detection and skew correction. In this sample of image the 53.81° angle is detected and the remaining images with skew corrected and displayed in different angles.

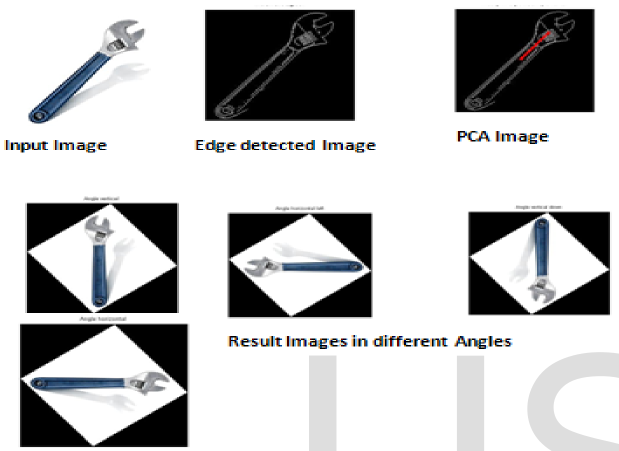

Figure-2: Sample result of ICDAR-2003

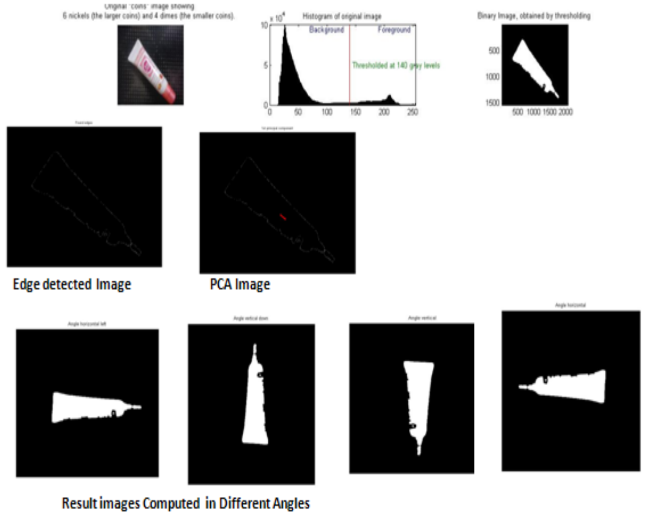
Original "coins" image showing
6 nickels (the larger coins) and 4 dimes (the smaller coins).
Histogram of original image
x 10^4
Background
Foreground
Thresholded at 140 grey levels
Binary Image, obtained by thresholding
Edge detected Image
PCA Image
Result images Computed in Different Angles

Figure-3: Sample result of our own database

5. Conclusions and Future work

In this paper we have proposed a method to detect and correct the skew in the image using PCA. Skew is inexorably introduced into the scanned document during scanning. So it is necessary to detect and correct the skew of the object in the image. PCA is applied to the edge image and skew of the object in the image is detected. Detected skew is used to manipulate the angle of the object in the image. Extensive experiments are carried out on standard skewd images database available on internet database and our own data set. Experiments show the good performance of the algorithm and the accuracy of the estimated angle is very high and the results are satisfactory and are more competitive.

Further this work may be extended to strengthen the performance of the algorithm to get the higher detection rate with more skewed angles with very complex background.

Acknowledgements:

Authors would like to thank Dr. Shivakumara, Palaiahnakote, University of Malaya, Faculty of Computer Science and Information Technology, Kuala Lumpur, Malaysia for critical comments and valuable suggestions

References

[1] D. Kumar et. al., "Modification Approach of Hough Transform for Skew Detection and Correction in Documented images", *International Journal of Research in Computer Science*, Vol.2. 2012.

[2] Reza Safabakhsh et. al., "Document Skew Detection Using Minimum Area Bounding Rectangle" . *Department of Computer Engineering Amirkabir University of Technology Tehran, Iran.*

[3] Sepideh Barekat Rezaei et. al., "Skew Detection Of Scanned Document Images", *Proceedings of the International Multi-Conference of Engineers and computer Scientists* 2013,Vol I, IMECS 2013,Mar 13-15,2013, Honk Kong .

[4] Joost Van Beusekomb, Faisal Shafaita, Thomas M. Breuela (2009)" Resolution Independent Skew and Orientation Detection for Document Images", *Image Understanding and Pattern Recognition (IUPR) Research Group.*

[5] Manjunath Aradhya V N et. al., "Skew Detection Technique for Binary Document Images based on Hough Transform", *International Journal of Communication Engineering and Technology*, PP:493-499, 2007.

[6] S. Lu et. al, "Automatic document orientation detection and categorization through document Vectorization", *in MULTIMEDIA, Proceedings of the 14th annual ACM international conference on Multimedia, ACM,* (New York, NY, USA), 2006.

[7] Khairuddin Omar et al, "Skew Detection and Correction technique For Arabic Document Images Based on Centre Of Gravity", *Journal of Computer Science*- 5(5):363-368, ISSN 1549-3646, 2009.

[8] Zhixin Shi and Venu Govindaraju" Skew Detection for Complex Document Images Using Fuzzy Runlength" *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)* 0-7695-1960-1/03 $17.00 © 2003 IEEE.

[9] P Shivakumara et al "A novel technique for estimation of skew in binary text document images based on linear regression analysis" *Journal of S¯adhan¯a* Vol. 30, Part 1, February 2005, pp. 69–85. © Printed in India.

[10] A. Alaei et. al., "A Painting Based Technique for Skew Estimation of Scanned documents", in Document Analysis and Recognition

(ICDAR), 2011 *International Conference on Document Analysis and Recognition, 2011.*

[11] Wassim Al-Khawand et.al., “A Novel Skew Estimation Approach Based on Same Height Grouping” *International Journal of Signal Processing, Image Processing and Pattern Recognition* Vol.7, No.3 (2014), pp.421-432 http://dx.doi.org/10.14257/ijsip.2014.7.3.34.

[12] Jonathan J. Hull, “Document Image Skew Detection: Survey and Annotated Bibliography Document Analysis Systems” *World Scientific,* pp. 40-64, 1998.

[13] Bishakha Jain et. al., “A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical Projection Profile Analysis”, *International Journal of Scientific and Research Publications,* Volume 4, Issue 6, June 2014 ISSN 2250-3153.

[14] Neha Watts et. al., “Performance Evaluation of Improved Skew Detection and Correction using FFT and Median Filtering”, *International Journal of Computer Applications* (0975 – 8887) Volume 100 – No.15, August 2014.

[15] P. Malathi, “ Skew Detection based on Bounding Edge approximation”, *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.

[16] Rangachar Kasturei et. al., “Document image analysis: A primer”, *Journal of S¯adhan¯a* Vol. 27, Part 1, February 2002, pp. 3–22. © Printed in India.

[17] Ruby Singh et.al., “Skew Detection In Image Processing”, *International Journal of Computer Technology & Applications*, Vol. 4(3),478-485, ISSN:2229-6093, May-June 2013.

[18] Lindsay I Smith, “A tutorial on Principal Components Analysis”, February 26, 2002.

[19] Jonathon Shlens, “A Tutorial on Principal Component Analysis: Mountain View”, CA 94043, Revised Dated: April 7, 2014; Version 3.02.

[20]