

Design and Implementation Recognition System for Handwritten Hindi/Marathi Document

Rupal Saoji¹, Prof. S. W. Mohod², Prof. M. C. Nikose³

¹ M.Tech. Computer Science & Engineering

² Co-ordinator, Department of Computer Engineering, BDCE Sevagram

³ Prof., Department of Computer Engineering, BDCE Sevagram

¹ rupal.saoji@gmail.com, ² sudhir_mohod@rediffmail.com, ³ monalinikose18@gmail.com

Abstract: In the present scenario most of the importance is given for the “paperless office” there by more and more communication and storage of documents is performed digitally. Documents and files which are present in Hindi and Marathi languages that were once stored physically on paper are now being converted into electronic form in order to facilitate quicker additions, searches, and modifications, as well as to prolong the life of such records. Because of this, there is a great demand of such software, which automatically extracts, analyze, recognize and store information from physical documents for later retrieval. Skew detection is used for text line position determination in Digitized documents, automated page orientation, and skew angle detection for binary document images, skew detection in handwritten scripts, in compensation for Internet audio applications and in the correction of scanned documents.

I. Introduction

In the present scenario most of the importance is given for the “paperless office” there by more and more communication and storage of documents is performed digitally. Documents and files that were once stored physically on paper are now being converted into electronic form in order to get quicker additions, searches, and modifications, as well as to prolong the life of such records. Because of this, there is a great demand for software, which automatically extracts, analyze, recognize and store information from physical documents for later retrieval.

In document image processing, skew angle detection is a very important part in data processing and it's the foundation of image analysis and recognition. In image based digital identification systems, the reliability of recognition is closely related to the quality of image data. Therefore, in most real-time document image processing, skew angle should be confirmed quickly and accurately to improve the accuracy of collection and entry for document information, mean while, to reduce the rejection rate and improve reliability and adaptability of systems. Most scanners have the capabilities of automatic de-skew which can segment the skew document images from the background, However, skew is often happened due to print in practice, the result is that the images can't be de-skewed correctly. Therefore, the research about de-skew algorithm content-based of document images could better reflect the nature of the problem and have a great significance in document image processing.

The different types of skews within a document page can fall into these categories:-

- **Global Skew:** - assuming that all page text has the same skew angle.
- **Multiple-Skew:**-when certain area of the page has one angle and other have different angle.
- **Non- Uniform text line skew:** - when the skew angle is fluctuates in the page.

Skew detection is used for text line position determination in Digitized documents, automated page orientation, and skew angle detection for binary document images, skew detection in handwritten scripts and in the correction of scanned documents. The largest classes of methods for skew detection are available which based on projection profile analysis, Hough transform, nearest-neighbor clustering, Fourier- transformation, histogram analysis and Neural Network.

In order to improve the readability and the automatic and quick recognition of handwritten Hindi and Marathi document images, preprocessing steps are imperative. These steps in addition to conventional steps of noise removal and filtering include text normalization such as baseline correction, slant normalization and skew correction and detection. These steps make the feature extraction process more reliable and effective.

II. Related Work

Different methodologies for the text skew estimation in binary and gray scale images. They have been used widely for the skew identification of the printed text. Some of available algorithms provide better accuracy but they are slow in speed, others have angle limitation drawback. So a new technique for skew detection in the paper, will reduce

the time and cost. Another method for skew detection and correction of handwritten Devanagari script using Hough transformation it works at word level which is applied on each word for skew detection and rotation transformation is used for correcting that skew. If skew of the word is in positive angle then it is corrected by rotating it in clockwise and in anti-clockwise direction. Skew detection and correction methods are used to align the handwritten text document by making the rectangular shape such as paragraph, text lines and tables. This new technique for skew detection will reduce the time and cost. Handwritten recognition has been one of the active and challenging task. Handwritten characters are written in various curved & cursive ways with different sizes, orientations, thickness and dimensions which is most difficult task to recognize such handwritten characters by any machine. In this paper a novel technique for Handwritten Hindi Character Recognition System is implemented using Canny Edge Detection technique and artificial neural network. The estimation and correction of skew is a difficult and challenging task in handwritten word since it has to be independent of the variations due to handwriting style and writing conditions of different persons. In this paper, a coarse-to-fine technique that integrates core-region information is presented. At first, a rough estimation and correction of the skew is accomplished by cutting vertically for detecting the center of mass in each part and calculating the inclination of the line that connects the two centers of mass of the overlapping area. To detect the core-region of the word (ascenders and descenders). The inclination of the line that connects the updated centers of mass corresponds to a finer estimation of the word skew. After correcting the detected skew the last step of core-region detection and skew correction is repeated iteratively. It have proposed schemes for skew detection and correction, segmentation of handwritten Kannada document using bounding box technique, Hough transform and contour detection respectively. Detecting text line is still a challenge to segment handwritten text lines in the general sense given no prior knowledge of script. image segmentation problem by enhancing text line structures using a Gaussian window and adopting the level set method to evolve text line boundaries. Variations in inter-line gaps and skewed or curled text-lines are some of the challenging issues in segmentation of handwritten text-lines. Moreover, overlapping and touching text- lines that frequently appear in unconstrained handwritten text documents increase segmentation complexities.

III. Proposed Work

The proposed methodology of the project are includes 5 steps which are input document, pre-processing, segmentation, skew detection and at last skew correction. So

that we can recognize scanned handwritten text document accurately.

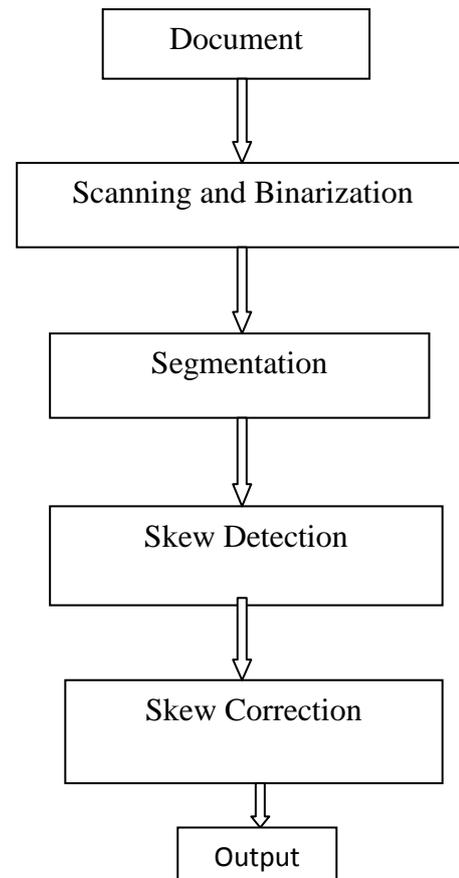


Fig: 1 Proposed System

A. Input Image

In which we can take that handwritten text document of Hindi and Marathi languages. Then by scanner or by camera we can get scanning image of that text documents. An image is a two-dimensional function $f(x,y)$.

B. Preprocessing

Binarization is simply extract the foreground data from background data. Digital Document means scanning or photography of the document. Normally digital document in the form of RGB format. A colour image can be converted to gray scale image. Then gray scale can be converted to binary image. Therefore with binary numbers it is easy compute and recognizes the characters.. To convert to Binarization we may use Thresholding.

Improper scanning of document or malfunction of camera or low quality of the documents , the digital document added by noise. It disturbance the recognition of the characters and numerals. . We can use Sauvola's algorithm to the Binarization of the gray scale image. The gray level image is back and white pixels. Then classify all the pixels on the basis

of thresholding values. The pixels value above the threshold is in white and value below the threshold are in black

The threshold is computing using Sauvola's Algorithm.

$$T=M*[1+K*(S/R-1)]$$

K- user defined parameter

M, S- mean and local standard derivation computed in windows size W centered on current pixel(size is user defined)

R- dynamic range of standard derivation(R=128 with 8-bit gray level)

C. Segmentation

Segmentation is the process of partitioning a digital text image into multiple segments. The goal of segmentation is to simplify or change the representation of an image into something that is more meaningful and easier to analyze.

Line Segmentation

In handwritten text document line of the text are segmented from the each other. this process is takes place by using line segmentation.

Word segmentation

In such type of segmentation the words are get segmented from the line of the handwritten text document.

Character Segmentation

The single character is get segmented from the word of the handwritten text document.

Bounding Box

Bounding Boxes [4] for connected components are the properties of the labeled connected component regions. A bounding box of a labeled region is a rectangle that just encloses the region completely. When a specific bounding box is determined for a connected region, the co-ordinates of the corners of the bounding box and its width and height are available. A bounding box completely specifies the boundaries of the corresponding connected component. In our method, we use filled bounding boxes which completely cover the corresponding connected components.

Min cut max flow Algorithm

energy minimization problems can be reduced to instances of the maximum flow problem in a graph

The maximum value of an s-t flow is equal to the minimum weight of an s-t cut.

The minimal cut of this graph will cut all the edges connecting the pixels of different objects with each other.

If f is a flow, then the net flow across the cut (S, T) is defined to be $f(S, T)$, which is the sum of all edge capacities from S to T subtracted by the sum of all edge capacities from T to S

The capacity of the cut (S, T) is $c(S, T)$, which is the sum of the capacities of all edge from S to T

A minimum cut is a cut whose capacity is the minimum over all cuts of G .

D. Skew Detection

Detection there are some methods are presents which are Hough transformation, nearest-neighbor clustering, Fourier-transformation, neural network. This angle in between 1-3 degree the it get neglected and the angle is above the 5 degree are go to next stage that is skew correction. The skew angle of each text line has been calculated. But the length of the text line may be very different. The longer the length of a text line is, the more truly reflected the skew angle of the image is. So it is not good to simply calculate the skew angle of each objective and average them. Using the length of text lines objectives as the weight is a reasonable calculation method. longer the length of a text line is the greater the weight will be.

We can get the skew angle by using formula

$$\cos(s) = l/m$$

$$s = 1 \setminus \cos(l/m)$$

where l – x coordinat of the pixel

m - distances of pixel from origin

Using the largest text lines characteristic as the skew angle of a document image is reasonable. The longer the length of a text line is, the higher the accuracy is.

E. Skew Correction

After the detection of skew angle this angle gets corrected by using the skew correction technique. So, we can use pixel orientation technique to do the skew correction because, this technique is concentrated on the pixel.

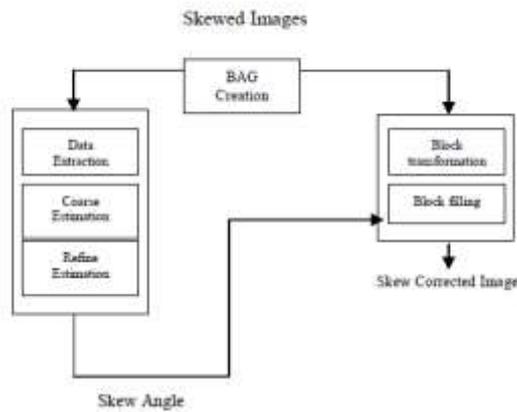


Fig 2:- Skew Correction

This method uses a coarse/refine strategy based on Hough transformation of connected components in the image. Speed and accuracy has been achieved as this confine the region of interest to a small and plausible text area, and confine the search range of Hough transform to Q , where Q is a coarse skew angle. considers some selected characters of the text which may be subjected to thinning and Hough transform to estimate skew angle accurately.

The method has two stages. In the first stage, selected characters from the document image are blocked and thinning is performed over the blocked region. In the second stage, the thinned coordinates are fed to Hough transform (HT) to estimate the skew angle accurately. The block diagram of the proposed methodology

IV. Result



Fig Input Image

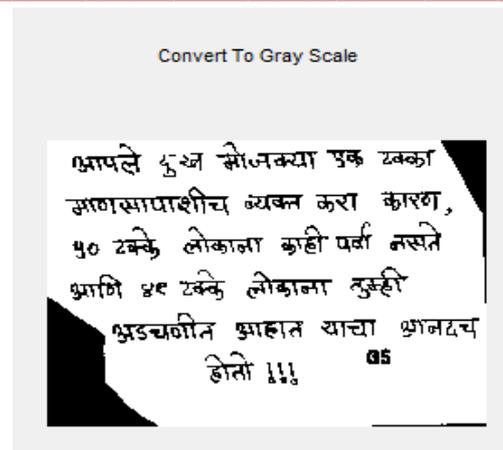


Fig 4: Binarized Image

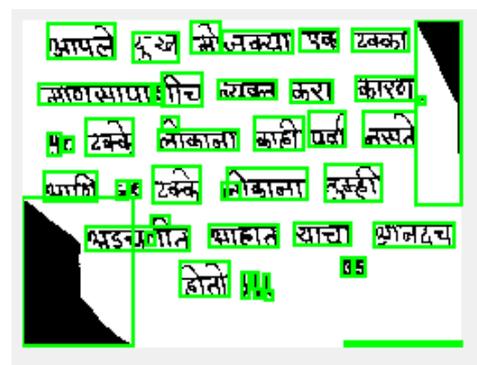


Fig 5: Segmented



Fig 6: Result

V. Conclusion

As you all know that handwritten documents in Marathi and Hindi or any language are very difficult to maintain an to

store for life long. So it is more important to convert that handwritten text document in digitalize form. So that we can easily perform operation on that handwritten document. To do digitalization we can use the scanning, binarization, segmentation, skew detection, skew correction processes. It gives accuracy and reliability in skew angel detection an correction process.

References

- [1] Ruby singh 1, Ramandeep Kaur 2, "Skew Detection in Image Processing", International Journal computer Technology and Application, Vol 4(3),478-485 IJCTA May-2013.
- [2] Trupti A. Jundale, Ravindra S. Hegadi," Skew Detection and Correction of Devanagari Script Using Hough Transform", International Conference on Advanced Computing Technologies and Application, ICACTA 2015.
- [3] Ambica Rani, Er. Harinderpal Singh, " A Review On Various techniques for Skew Detection And Correction In Handwritten Text Documents", IJREAT International Journal Of Research in Engineering And Advance technology, Volume 3, Issue 3, June-July 2015.
- [4] Tanuja K. , Usha Kumari V. And Sushma T. M. ," Handwritten Hindi Character Recognition System Using Edge detection and Neural Network", International Journal of Advanced Technology and Engineering Exploration ISSN, Volume 2, Issue 6, May 2015.
- [5] A. Papandreou and B. Gatos, "A coarse to Find Skew Estimation Technique for Handwritten Words", 2013 12th international Conference on Document Analysis and Recognition.
- [6] Mamatha Hosalli Ramappal and Shrikantamurthy Krishnamurthy, "Skew Detection, Correction and Segmentation of Handwritten Kannada Document", International Journal of Advanced Science and Technology vol . 48, Nov 2012.
- [7] Panwar, Subhash, and Neeta Nain. "A Novel Approach of Skew Normalization for Handwritten Text Lines and Words." Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on. IEEE, 2012.
- [8] Manjunath , G. Hemantha, and P. Shivakumara., "Skew Detection technique for Binary Document Images based on Hough Transform", 2007.
- [9] Li, Yi, Yefeng Zheng, and David Doermann. "Detecting text lines in handwritten documents." Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Vol. 2. IEEE, 2006.
- [10] P. Shivakumara, G. H. Kumar, D. S. Guru and P. Nagabhushan, "A novel technique for estimation of skew in binary text document images based on linear regression analysis", Sadhana, vol. 30, Part 1, (2005) February, pp. 69-85.
- [11] V. N. M. Aradhya, "Document Skew Estimation: An Approach based on Wavelets", ICCCS'11, (2011), pp. 359-364.