© 2016 IJSRSET | Volume 2 | Issue 4 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Skew Detection Techniques Used In Scanned Document Images

Snehal S. Kolhe, K. T. Jadhao

EXTC Department, Mumbai University, Maharastra, India

ABSTRACT

In this paper skew detection technique for printed and handwritten devnagri scanned documents is proposed. Recognition methods for printed and handwritten texts in Scanned documents are significantly different. Skew detection technique consists of processing steps like preprocessing, segmentation, skew correction and detection with enhancement which is used to identify writer. Segmentation is used to extract text lines and words from handwritten and printed documents. A horizontal projection and vertical projection algorithm is used to segment the document into a number of lines. Nilback, Sauvola's and wolfs algorithm is used for binarization. Linear Otsu thresholding is used for filter. The scan line method is used for the skew detection and correction .Adaptive histogram algorithm is used for image enhancement. Various Printed and handwritten text documents are scanned for the different preprocessing methods. In segmentation, for the Line, word, character segmentation we used the separate algorithms. For the skew detection and correction, skew line method gives the better accuracy.

Keywords: Pre-Processing, Filtering, Segmentation, Skew Detection and Correction, Enhancement

I. INTRODUCTION

Invention of new technologies leads us towards the achievement of paperless office and paperless society. Digital script analysis is first step of automating the offices. Everywhere documents are there in the form of papers like forms, check etc.Most of the documents are both handwritten and printed or combination of both. For example, railway reservation forms, bank cheques. Handwritten and printed text is most of the times interlaced at word, line and character level. So the identification of such documents is a challenging task. Separation of handwritten and printed text from such documents is very essential. Using segmentation method we can separate document at word, Character and line level which will reduce the search time and avoids the confusion.

Document scanning is the first step in textual processing. During document scanning, skew is introduced into the document image. Skew affects the recognition of text document. The skew angle detection algorithm extracts the straight lines correctly, and increases the reliability of the estimate angle. Document images are more complex. There are many differences in language writing also the differences of printed font and handwriting font, and the text size, colour also have a larger difference in different application areas.

Skew is the text which neither parallel nor at right angles to a specified or implied line. Skew angle detection and correction in scanned digital document images is very difficult and critical step during layout analysis. For text line position determination in Digitized documents, skew detection is used. Skew angle correction makes automated orientation of the page for deskewing image.

II. METHODS AND MATERIAL

A. Literature Survey

Document image segmentation to text lines and words is a critical stage towards unconstrained handwritten document recognition. Variation of the skew angle between text lines or along the same text line, existence of overlapping or touching lines, variable character size and non-Manhattan layout are the challenges of text line extraction. Due to high variability of writing styles, scripts, etc., methods that do not use any prior knowledge and adapt to the properties of the document image, as the proposed, would be more robust. Line extraction techniques may be categorized as projection based, grouping, smearing and Hough-based. [11]. This paper is implementation of preprocessing steps for the handwritten and printed text document .As the name suggest, implementation of preprocessing methodologies mean it have the number of preprocessing methods are use and chose the best method with less time with good results Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Hence, preprocessing is the preliminary step which transforms the data into a format that will be more easily and effectively processed. Therefore, the main task in preprocessing the captured data is to decrease the variation that causes a reduction in the recognition rate and increases the complexities.[4] The main objective of the preprocessing stage is to normalize and remove variations in the handwritten and printed text document. In character recognition systems most of the applications use grey or binary images, since processing colour images is computationally high. Such images may also contain non-uniform background and/or watermarks making it difficult to extract the document text from the image without performing some kind of preprocessing. Few or some of these techniques or others may be used at different stages of the OCR system by Dan WANG, Xichang WANG in A Skew Angle Detection Algorithm based on Maximum Gradient Difference in 2011. [12] In 2009, Vassilis Papavassilioua, Themos Stafylakisa, Vassilis Katsourosa proposed Handwritten document image segmentation into text lines and words, in that the line segmentation algorithm is based on locating the optimal succession of text and gap areas within vertical zones by applying Viterbi algorithm. Then, a text-line separator drawing technique is applied and finally the connected components are assigned to text lines. Word segmentation is based on a gap metric that exploits the objective function of a soft-margin linear SVM that separates successive connected components. The algorithms tested on the benchmarking datasets of ICDAR07 handwriting segmentation contest and outperformed the participating [15] In 2011, QI Xiaorui, MA Lei, SUN Chang jingo and LIU Jiang, proposed Fast Skew Angle Detection Algorithm for Scanned Document Images based on 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, meanwhile, 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. [13]

In 2012, Upasana Patil1, Masarath Begum proposed, word level handwritten and printed text separation based on shaped features. In this method the discriminating handwritten and printed text from document images based on shape features. The separation of handwritten and printed text from document image is essential to optimize the OCR accuracy and to activate an appropriate OCR engine. It leads to reduce the search space of the OCR and it also facilitates the retrieval of Handwritten and Printed text from document images. The used of IAM dataset 3.0 and with morphological transformations segmented 74 pages and obtained 10768 words and 2000 were used for experimentation and achieved average accuracy of 98.57% with only seven features. The proposed method is simple, have promising discrimination accuracy and less time complexity. [11] In 2013, Ruby Singh, proposed Skew detection in image processing many researchers proposed different methodologies for the text skew estimation in binary images/gray scale images. They have been used widely for the skew identification of the printed text. There exist so many ways algorithms for detecting and correcting a slant or skew in a given document or image. Some of them provide better accuracy 70% but 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. [2] In 2013, Utkarsh Mathur, Rohit Sharma, Naveen Srivastava proposed, script independent angular skew Detection and correction algorithms, document digitalization is done in a large scale for hand written and printed documents, where the documents are scanned and stored in a digital form. Maximum numbers of these documents are hand written, and hence, they comprise of various errors like angular skew of words, or sentence as a whole. Before the document can be digitalized using OCR software, it is pre-processed for angular skew detection and their removal from the scanned document. Detection and correction of angular skewness of the words is the most difficult task for the process, which must be carried out as this skewness is present in almost all of the hand written words. [7]

B. Development of System

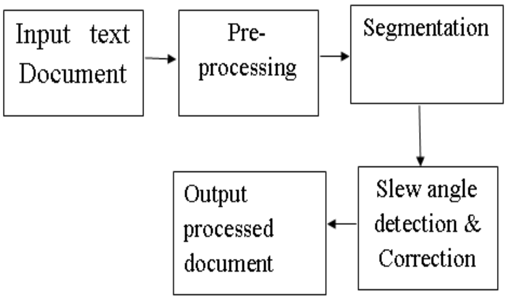


Figure 1. System Diagram

Preprocessing steps includes binarization, filteration, skew angle detection and correction. In binarization, Niblack method of gray scale binarization is used. Linear otsu filter is used to filter noise from the image. As compared to scanners, cameras offer fast, flexible and non-contact document imaging, but with distortions like uneven shading and warped shape. So, camera-captured and scanned document images need preprocessing steps. These steps includes steps binarization,filteration,skew detection, segmentation.
These pre-processing steps can be applied on the scanned document or camera captured images.

3.1 Image acquisition

Image acquisition is used to obtain the image of document in colour, gray level or binary format which is captured with camera or scanner.

3.2Image pre-processing

Image preprocessing steps involves binarization, filteration, segmentation, skew detection.

3.2.1 Binarization

Image binarization is to select a threshold value, and classify all pixels with values above threshold as white, and remaining pixels as black. Selecting proper threshold is very important task. In many cases, finding one threshold computing to the entire image is very difficult, and in many cases even impossible. Binarization is processing of converting colour image in to binary image. There are various Binarization methods and in that various different algorithm such as Nilback, Sauvola, Wolfs, Otsu etc used for the exact and accurate output. By comparing all algorithms used the Nilback algorithm is good for the binarization of the document for the better result.

Niblack Algorithm

The calculation of threshold value is depending on the local mean m and the standard deviation s of all the pixels in the window and threshold is given by formula

$$T_{niblack} = m + k\sqrt{\sum \frac{P_i^2 - m^2}{N_p}}$$

- Where NR_p*R* is the number of pixel in the Pi gray image
- m is the average value of the pixels and k= -0.2

Steps for Nilback Algorithm:

1. The input image is successively sub sampled to different scales.
2. Calculate the local mean value for the current pixels
3. Calculate the local standard deviation for the neighbour pixels.
4. From both these calculates the threshold for different part of the image.
5. From the threshold image, the input image is binarized.

3.2.2 Filtering

During data extraction process binarization is required sometimes. Binarization sometimes discards the image. It replaces the pixels in the background with binary 0 and 1 in the image. Filters are mainly used to suppress either the high frequencies in the image. Linear filtering reduces the noise; it sharpens the edges by filtering the image by correlation with an appropriate filter.

Steps for Linear Otsu thresholding Algorithm:

1. Compute histogram and probabilities of each intensity level

2. Set up initial X and Y
3. Step through all possible thresholds t=1 maximum intensity
    1. Update x and y
    2. Compute w
4. Desired threshold corresponds to the maximum W
5. You can compute two maxima (and two corresponding thresholds). P is the greater max and Q is the greater or equal maximum
6. Desired threshold = $\frac{P + Q}{2}$

3.3 Segmentation

Segmentation is the process of partitioning a digital text image into multiple segments. Segmentation of document into lines, words and characters is critical for handwritten document recognition. Line extraction techniques may be categorized as projection based, grouping, smearing and Hough-based. [11].

Using text line segmentation word segmentation grouping of text lines into paragraphs, characterization of text lines as titles, headings, footnotes, etc. tasks can be developed. Text line segmentation is a critical stage in layout analysis, upon which further tasks such as word segmentation, Segmentation simplifies the image or changes the representation of an image into meaningful and easier to recognise. Segmentation can be of three type text, Line segmentation, Word segmentation, Character segmentation.

3.3.1 Line Segmentation

Line segmentation separates line from the text document.

Steps for line segmentation algorithm:

1. Use the horizontal projection method for segmenting lines from text.
2. Count the white pixel in each row.
3. Find minimum and maximum values of the rows
4. Find minimum and maximum values of the columns
5. The values of rows and columns give no white pixels
6. Replace all such rows and columns by 1
7. Invert the image to make empty rows as 0 and text lines will have original pixels.
8. Crop the line from the min and max values of rows and columns.

3.3.2 Word Segmentation

Word segmentation is to separate word from the line.

Steps for word segmentation Algorithm

1. Label and count connected components
2. Use the vertical projection method for segmenting word from each line.
3. Count the white pixel in each row.
4. Find minimum and maximum values of the rows
5. Find minimum and maximum values of the columns
6. The values of rows and columns give no white pixels
7. Replace all such rows and columns by 1
8. Invert the image so that text lines will have original pixels.
9. Crop the word.
10. Save the word in the file.

3.3.3 Character Segmentation

Character Segmentation separates character from the word.

Steps for character segmentation Algorithm

1. Label and count connected components
2. Use the vertical projection method for segmenting characters from word.
3. Count the white pixel in each row.
4. Find minimum and maximum values of the rows
5. Find minimum and maximum values of the columns
6. The values of rows and columns give no white pixels
7. Replace all such rows and columns by 1
8. Invert the image to make empty rows as 0 and text lines will have original pixels.
9. Crop the character.
10. Save the characters in the file.

3.4 Skew detection and correction

Skew angle Detection and correction is a very important part in data processing and it is the foundation of image analysis and recognition. Document skew is a distortion that often occurs during document scanning or copying. Skew is the orientation of text lines. Skew can also be intentionally designed to emphasize important details in a document.

When the skew of the document image is zero degrees, the projection profile peak times will be longer. To understand the reason, consider the scan lines are drawn on two document image with 5 and zero degrees of skew. Here, the scan line means a row of the image. Scan lines plotted on document image with 5 degrees of skew, include white areas between text lines, and while most of scan lines plotted on document image with zero degrees of skew, include a text line and no white areas between text lines. So, in those rows of the image with zero degrees of skew, the number of black pixels is higher and the projection profile peaks are longer. Therefore, the projection profile can be used as a suitable feature for skew detection. It is needed to create a feature to describe which one is more peaked for comparing peaks of projection profiles. So employing a criterion function provides a numerical description of the peaks. [8] Thus the scan line method for the skew detection and correction is as follows.

Mathematical Analysis

The text lines' starting point and ending point of the objective marked with "t" are (Xs, Ys), (Xe, Ye), and then the skew angle of the text lines can be estimated as,
$\Theta t = (Ye - Ys) / (Xe - Xs)$ ------------------- (7)

Defining the angle energy of text lines, assuming the skew angle of the objective marked with t is t, Ѳt, and the length of a text line (the number of the objective pixels) is Ct. and then the angle energy of the text lines is:

$Pt = \Theta t * Ct$------------------------ (8)

There are M text lines characteristics of a document image. Then the skew angle of the document Ct image is eventually defined as:

$\theta = \sum_{t=0}M-1\ Pt/\sum_{t=0}M-1$ --------------- (9)

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. [2]

If $\theta<0$,

$\theta=\theta+180$……………………… (10)
Or
if $\theta>85$,
Rotate the image with its correct angle.

Steps for skew detection and correction Algorithm

1. Select the input image.
2. Convert to grey scale image.
3. Apply the Horizontal gradient method.
4. Calculate text line character.
5. Calculate text line objective tracking.
6. Calculate the skew angle for the input image.
7. Apply the block transformation with filling to correct the skew.
8. If the skew angle is >85 then it will rotate the image with the rotated angle.
9. Display Skew corrected image.

3.5 Image Enhancement

Image enhancement is widely used in the field of image processing where the subjective quality of human being is important to human interpretation aim of the image enhancement is to provide `better' input for other automated image processing.

Enhancement is the improvement of an image to alter impact on the viewer. In this paper. For image enhancement adaptive histogram equalization method is used.

Steps for adaptive histogram thresholding Algorithm

1. Obtain all the inputs: Image, Number of regions in row and column directions, Number of bins for the histograms used in building image transform function, normalized from 0 to 1.
2. Pre-process the inputs: Determine real clip limit from the normalized value if necessary, pad the image before splitting it into regions.
3. Process each contextual region thus producing gray level mappings: Extract a single image region, make a histogram for this region using the specified number of bins, clip the histogram using clip limit, and create a mapping for this region.
4. Interpolate gray level mappings in order to

assemble final image: Extract cluster of four neighbouring mapping functions, process image region partly overlapping each of the mapping tiles, extract a single pixel, apply four mappings to that pixel, and interpolate between the results to obtain the output pixel; repeat over the entire image.

III. RESULTS AND DISCUSSION

Scan line method can detect and correct the skew for scanned document .Handwritten document segmentation into line, word; character is possible using horizontal projection and vertical projection method. Results for the methods used are mentioned in table no 1, 2 and 3.

Table1: Accuracy for handwritten text document

Total no of sample document images	%accuracy for line segmentation	%accuracy for word segmentation	%accuracy for character segmentation
10	96	90	80

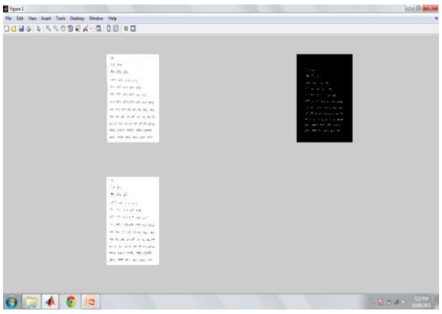
Table 2: Accuracy for Printed text document

Total no of sample document images	%accuracy for line segmentation	%accuracy for word segmentation	%accuracy For character segmentation
10	95	90	77

Table 3: Error calculation for text document

Sr. No.	Original skew	Corrected skew	Error
1	4.00	3.73	0.27
2	4.12	3.16	0.96
3	-5.00	-3.77	1.23
4	-3.00	-2.24	0.76
5	4.00	3.68	0.32

FIGURES



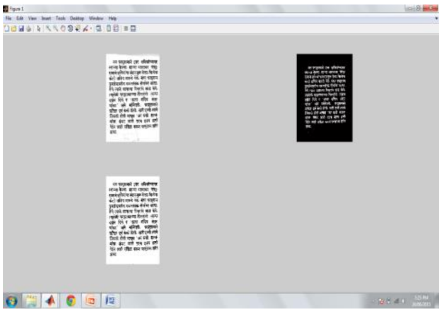
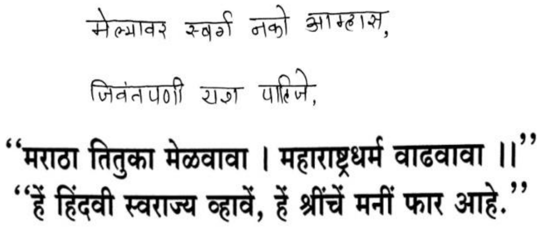


Figure 2. Output for the binarization with filtration for handwritten and printed document.



मेल्यावर स्वर्ग नको आम्हास,

जिवंतपणी राज्य पाहिजे,

“मराठा तितुका मेळवावा । महाराष्ट्रधर्म वाढवावा ।।”

“हें हिंदवी स्वराज्य व्हावें, हें श्रींचें मनीं फार आहे.”

Figure 3. Example for segmentation in handwritten and printed document.

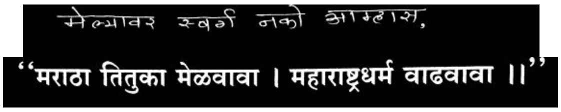

मेल्यावर स्वर्ग नको आम्हास,
“मराठा तितुका मेळवावा । महाराष्ट्रधर्म वाढवावा ।।”

Figure 4. Line segmentation for handwritten and printed document.

.

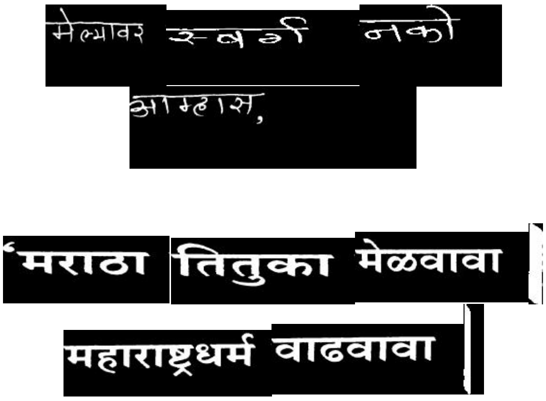


Figure 5. Word segmentation for printed and handwritten document

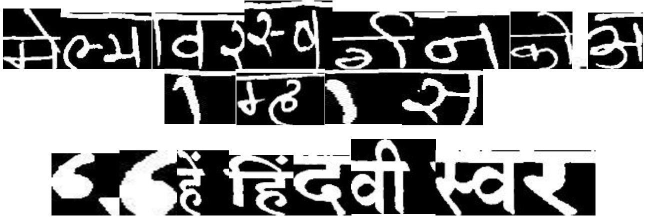


Figure 6. Character segmentation for handwritten and printed document.

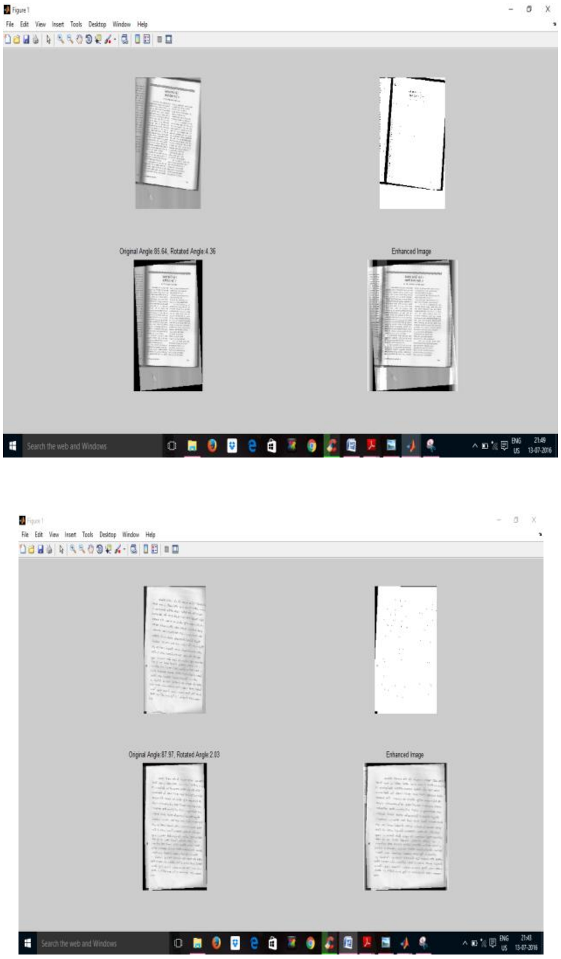

Original Angle:85.64, Rotated Angle:4.36
Enhanced Image
Original Angle:87.97, Rotated Angle:2.03
Enhanced Image

Figure 7. Skew detection and correction with enhancement in handwritten and printed document.

IV. REFERENCES

[1] P. Mathivanan, B. Ganesamoorthy and P. Maran "**Watershed algorithm based segmentation for handwritten text identification**" Ictact journal on image and video processing, february 2014.

[2] Ruby singh1, Ramandeep kaur2, "**Skew detection in image processing**", Int.J.Computer technology applications.vol 4 IJCTA | May-June 2013.

[3] NamishaModi,Khushneet Jindal "**Text Line detection and Segmentation in Handwritten Gurumukhi Scripts**" Volume 3, Issue 5, May 2013.

[4] MortezaValizadeh,Ehsanollahkabir,"**An adaptive water flow model for binarization of degraded document images**", IJDAR (2013).

[5] A. Papandreou and B. Gatos "**A Coarse Fine Skew Estimation Technique for Handwritten Words**" 1520-5363 2013.

[6] Irfan Ahmad "**A technique for skew detection of printed Arabic documents**" 978-0-7695-5051-0/13 $26.00 © 2013 IEEE DOI 10.1109/CGIV.2013.21 2013.

[7] Utkarsh Mathur, Rohit Sharma, Naveen Srivastava "**Script independent angular skew Detection and correction algorithms**" 978-1 - 4799-1 607-8/1 3/$31 .00©201 3 IEEE 2013.

[8] FirasAjilJassim, Fawzi H. Altaani, "**Hybridization of Otsu Method and Median Filter for Color Image Segmentation**" IJSCE) ISSN: 2231-2307, Volume-3, Issue-2, May 2013.

[9] SepidehBarekatRezaei,AbdolhosseinSarrafzadeh, and JamshidShanbehzadeh,**" Skew Detection of Scanned Document Images" Vol I,** IMECS 2013,

[10] Er.Mandeep kaur,Er.Kiran jain, "**Study if image enhancement techniques**" International Journal of advance research in computer science and software engineering vol3,Issue 4,April 2013 ISSN: 2277 128X .

[11] Upasana Patil1, Masarath Begum "**word level handwritten and printed text separation based on shaped features**" ISSN 2250-2459, Volume 2, Issue 4, April 2012.

[12] Dan WANG, Xichang WANG, Jiang LIU "**A Skew Angle Detection Algorithm based on Maximum Gradient Difference**" 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)

[13] Qi Xiaorui, Ma Lei, Sun Changjiang, Liu Jiang "**Fast skew angle detection algorithm for scanned images**" vol.32,no.21,2011.

[14] A S N Chakravarth, Penmetsa V Krishna Raja, Prof. P S Avadhani "**handwritten text image Authentication using back Propagation**" IJNSA, Vol.3, No.5, Sep 2011.

[15] VassilisPapavassilioua,ThemosStafylakisa,George Carayannis"**Handwritten document image segmentation into text lines and words**" Pattern Recognition 43