International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 4 422 - 426

Review of Skew Detection Techniques in Degraded Document Images

Snehal S. Kolhe
Department Of Electronics
And Telecommunication,
Armiet, Shahapur.
snehalk7@gmail.com

Prof. K. T. Jadhav
Department Of Electronics
And Telecommunication,
Armiet, Shahapur
ktjadhao@gmail.com

***Abstract*—**Now a day's document digitization is done in large scale for printed and handwritten documents where documents are scanned and stored in digitized form. Before the document isdigitalized using OCR software, it is pre-processed for angular skew detection and their removal from the scanned document.This paper makes study on skew angle detection and correction for printed and handwritten devnagri scanned document images. Skew angle detection is very important in data processing. It is foundation of image analysis and recognition. So in order to improve accuracy of collection and entry for document image skew angle should be confirmed quickly and accurately.

Keywords- *Pre-processing, Segmentation, skew detection and correction*

*****

I. INTRODUCTION

Keeping in mind the end goal to enhance the meaning fulness and the programmed acknowledgment of written by hand devanagri content archive, preprocessing steps are basic. These strides make the element extraction handle more solid and successful. In archive picture preparing, skew edge identification is a critical part in information handling and it's the establishment of picture examination and acknowledgment. In picture based computerized recognizable proof frameworks, the unwavering quality of acknowledgment is firmly identified with the nature of picture information. Consequently, in most constant report picture handling, skew point ought to be affirmed rapidly and precisely to enhance the exactness of accumulation and passage for archive data, in the mean time, to lessen the dismissal rate and enhance unwavering quality and flexibility of frameworks. Most scanners have the capacities of programmed de-skew which can portion the skew report pictures from the foundation, However, skew is frequently happened because of print practically speaking, the outcome is that the pictures can't be de-skewed accurately. Along these lines, the exploration about de-skew calculation content-based of report pictures could better mirror the way of the issue and have an extraordinary importance in archive picture handling.

Archives and documents that were once put away physically on paper are currently being changed over into electronic structure keeping in mind the end goal to encourage speedier augmentations, quests, and adjustments, and in addition to drag out the life of such records. Due to this, there is an incredible interest for programming, which consequently removes, break down, perceive and store data from physical records for later recovery. One of the imperative strides of report preparing is Textual handling through Optical character recognizer (OCR).

Skew alludes to the content which neither parallel nor at right edges to a predefined or suggested line. Character acknowledgment is exceptionally touchy to the page skew, skew discovery and amendment in record pictures are the basic strides before design examination. Skew discovery is utilized for content line position determination in Digitized reports, mechanized page introduction, and skew edge recognition for paired archive pictures, skew identification in written by hand scripts, in pay for Internet sound applications and in the rectification of checked records.

There are two stages for identifying the skew. The initial step is measurement diminishment; it is the procedure of diminishing the span of picture pixels of components and finding another element with lower measurements. It incorporates change of beginning set to other set for holding most extreme data. This will extricate the underlying set first and will produce missing data in it. After change, select an ideal subset of components in view of a goal capacity. There are distinctive measurement decrease techniques, which we can use in skew identification of report pictures. The best subset incorporates the greater part of the firmly pertinent and pitifully important however non-repetitive components. In the second step, skew is assessed. Here, the deviation relating to the most astounding and the least estimation of the capacity is generally considered as the skew.

The distinctive sorts of skews inside a report page can fall into these classifications:- Global Skew, expecting that all page content have the same edge skew, Multiple-Skew, when certain territory of the page have diverse inclination than other, and Non-Uniform content line skew, when the inclination varies. Skew discovery is utilized for content line position determination in Digitized records, mechanized page introduction, and skew edge identification for parallel archive pictures, skew recognition in written by hand devanagri scripts, in remuneration for Internet sound applications and in the revision of checked reports. The biggest classes of strategies for skew discovery depend on projection profile examination, Hough change, closest neighbor bunching, Fourier-change, histogram investigation, minutes and different techniques.

Machine reproduction of human capacities has been an exceptionally difficult exploration field following the approach of computerized PCs. In a few zones, which require certain measure of insight, for example, calculating or chess playing, enormous enhancements are accomplished. Then again, people still outflank even the most capable PCs in the generally routine capacities, for example, vision. Machine reenactment of human perusing is one of these ranges, which has been the subject of escalated examination throughout the previous three decades, yet it is still a long way from the last boondocks.

II. OVERVIEW

Preprocessing is one to alter the information either to right lacks in picture, or to set up the information for post handling of OCR. Information preprocessing portrays any kind of handling performed on crude information to set it up for another preparing technique. Subsequently, preprocessing is the preparatory step which changes the information into an arrangement that will be all the more effortlessly and adequately handled. Thusly, the principle errand in preprocessing the caught information is to diminish the variety that causes a lessening in the acknowledgment rate and builds the complexities. [4] The primary goal of the preprocessing stage is to standardize and evacuate varieties in the written by hand and printed content record. In character acknowledgment frameworks a large portion of the applications use dark or double pictures, since preparing shading pictures is computationally high. Such pictures may likewise contain non-uniform foundation and/or watermarks making it hard to remove the report content from the picture without performing some sort of preprocessing. To enhance OCR results, preprocessing of the picture is must. In the event that characters are thick and touching, help the shine. Few or some of these strategies or others might be utilized at various phases of the OCR framework by Dan WANG, Xichang WANG in A Skew Angle Detection Algorithm in view of Maximum Gradient Difference in 2011. [6] The binarization strategies reported in the writing are for the most part either worldwide or neighborhood. Worldwide techniques, utilizing a few criteria in view of the dim levels of the picture, locate a solitary edge esteem. These strategies look at the dark level of every pixel with a limit esteem and name it as either content or foundation. Worldwide techniques have great execution on account of good partition between the content and foundation dim levels , however when the histogram of the content covers with that of the foundation, these strategies neglect to work appropriately by Morteza Valizadeh , Ehsanollah Kabir in A versatile water stream model for binarization of debased report pictures in 2013.[7]

Numerous analysts proposed distinctive strategies for the content skew estimation in double pictures/dim scale pictures. They have been utilized broadly for the skew ID of the printed content. There exist such a large number of courses calculations for recognizing and amending an inclination or skew in a given report or picture. Some of them give better precision however are moderate in rate, others have edge restriction downside. So another procedure for skew location in the paper will diminish the time and cost in by Ruby Singh in Skew identification in picture preparing in 2013. [4]

III. DEVELOPMENT OF SYSTEM

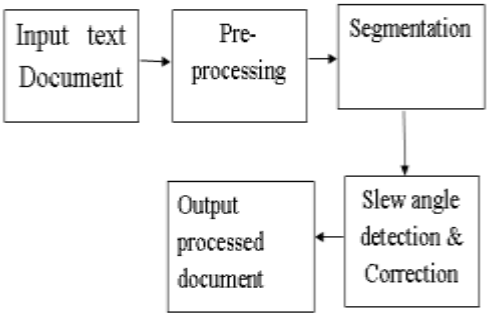


Fig. 1 System Diagram

SYSTEM DEVELOPMENT STAGES

1. INPUT TEXT DOCUMENT

Image acquisition is the input text document. Acquire image of any document with the help of camera or scanner. Image acquisition is used to obtain the image of document in color, gray level or binary format.

2. PRE-PROCESSING

These are the pre-processing steps often performed in OCR

1. Binarization

The least difficult approach to utilize picture binarization is to pick an edge esteem, and arrange all pixels with qualities over this edge as white, and every single other pixel as dark. Selecting legitimate edge is essential undertaking. Much of the time, discovering one limit good to the whole picture is extremely troublesome, and as a rule even outlandish. In this way, versatile picture binarization is required where an ideal limit is decided for every picture territory. Binarization is handling of changing over shading picture into parallel picture. In binarization, first we are changing over shading picture into Gray scale picture utilizing taking after recipe. [2]There are different Binerization techniques and in that different distinctive calculation utilized are as per the following. Shading picture is changed over into dim picture and taking after calculations are connected on dark scale picture for changing over it into twofold picture.

A. Niblack Algorithm

It is local thresholding algorithm. Local thresholding algorithms give good results for document because it calculate different threshold for different part of the image, considering pixel value. Niblack's algorithm calculates a pixel-wise threshold by sliding a rectangular window over the gray level image. 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.

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

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

B. Sauvola's algorithm:

In Sauvola's algorithm, Drawbacks of Niblack's method are removed. This is done by computing the threshold using the dynamic range of image gray-value standard deviation R. Threshold value for sauvola method is given by

$$T_{sauvola} = m * \left(1 - k * \left(1 - \frac{s}{R}\right)\right)$$

Where k is set to 0.5 and R is the threshold value to 128.
M is the average value.

C. Wolf's Algorithm:

In Wolf algorithm the mean gray value of the image is used in the threshold .Threshold is given as below.

$$T_{wolf} = (1-k) * m + k * M + \frac{(k * s)}{R(m - M)}$$

Where k is fixed to 0.5, M is the minimum gray value of the image and R is set to the maximum gray-value standard deviation obtained over all the local neighborhoods [2]

2. FILTERING

In addition to the removal of noise in documentary images. The data extraction procedure often requires binarizing the images, which discard most of the noise & replace the pixel in the image, character & the pixel in the background with binary 0 & 1 respectively. After binarization, document images are usually filtered to reduce noise. [6]

3. SEGMENTATION

Division is the procedure of apportioning a computerized content picture into various sections. The objective of division is to rearrange or change the representation of a picture into something that is more significant and simpler to break down. It will be isolated into three sections concerning the division of the content, Line division, Word division, Character division.

Line Segmentation:

It is only Separate line from the content archive.

Word division:

It is only separate word from the line.

Character division:

It is only separate character of the word.

There are different techniques in the divisions like,

Edge based division Histogram thresholding and cutting strategies are utilized to fragment the picture. They might be connected specifically to a picture, however can likewise be consolidated with pre-and post-preparing procedures.

Edge based division With this method, distinguished edges in a picture are expected to speak to question limits, and used to recognize these items.

District based division Where an edge based method may endeavor to discover the article limits and after that find the item itself by filling them in; an area based system takes the inverse methodology, by (e.g.) beginning amidst an article and afterward "developing" outward until it meets the item limits.

Edge location is an all around created field all alone inside picture preparing. Locale limits and edges are firmly related, subsequent to there is regularly a sharp change in force at the area limits. Edge identification procedures have in this way been utilized as the base of another division strategy. The edges recognized by edge recognition are regularly disengaged. To section an article from a picture anyway, one needs shut area limits. The fancied edges are the limits between such questions. Division is the procedure of partitioning picture into discrete areas and ordinarily goes before all picture examination. [2][1]

4. SKEW DETECTION AND CORRECTION

Report skew is a bending that regularly happens amid archive examining or duplicating. This primarily concerns the introduction of content lines and with no skew the lines are even or vertical, contingent upon the dialect. Skew can likewise be deliberately intended to underline vital points of interest in a record. Along these lines, this impact is unavoidable in numerous genuine cases and it ought to be disposed of on the grounds that it drastically decreases the precision of the resulting strategies, for example, page division/grouping and OCR. In spite of the fact that few techniques for page division and grouping without skew identification and redress are known they however either deliberately confines the noticeable point extent or they are considered in disengagement from the ensuing preparing operations, including character division and acknowledgment. [4]

Skew point Detection is a vital part in information handling and it is the establishment of picture examination and acknowledgment. In picture based Digital distinguishing proof frameworks, the dependability of Recognition is firmly identified with the nature of picture information. Thusly, in most ongoing record picture preparing, skew point ought to be affirmed rapidly and precisely to enhance the exactness of gathering and section for report Information, in the interim, to lessen the dismissal rate and Improve unwavering quality and versatility of frameworks. Most Scanners have the capacities of programmed de-skew which Can section the skew archive pictures from the Background, However, skew is frequently happened because of print by and by, the outcome is that the pictures can't be de-skewed Correctly. In this way, the exploration about de-skew calculation Content-based of record pictures

could better mirror the Nature of the issue and have an incredible essentialness in Document picture handling. The current techniques for skew edge estimation fundamentally have a few classifications: the strategy in view of Hough Transform, on cross-connection, on projection, on Fourier Transform and of K-Nearest Neighbor. The benefit of straight lines identification technique taking into account Hough change is not touchy to clamor, but rather the Amount of computation is bigger. In this way, the strategy ought to decrease the quantity of Hough change beyond what many would consider possible by and by, lessen the quantity of focuses that Participate in Hough change or diminish the skew edge Detection exactness. [2]

A. Scan line Method:

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, [7]

$\Theta t = (Ye - Ys) / (Xe - Xs)$

Defining the angle energy of text lines, assuming the skew angle of the objective marked with t is t, Θt, 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:

$Pt = \Theta t * Ct$

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

$$\theta = \sum_{t=0}^{M-1} p_t \Big/ \sum_{t=0}^{M-1} c_t$$

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. [7]

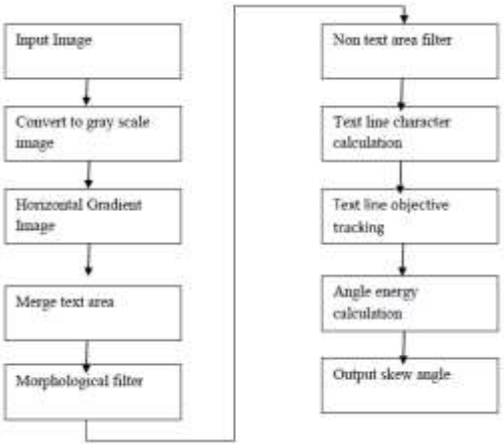


Fig. 2 Skew angle detection

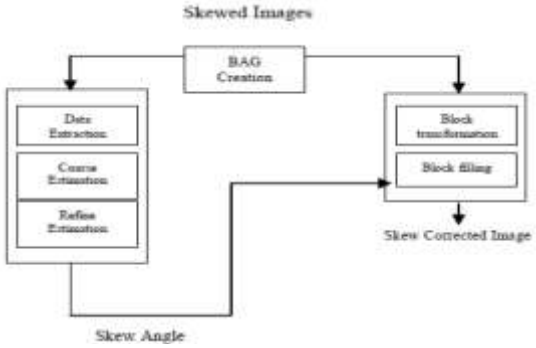


Fig. 3 skew angle correction

B. Hough Transform Method:

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 is given in Figure. [2]

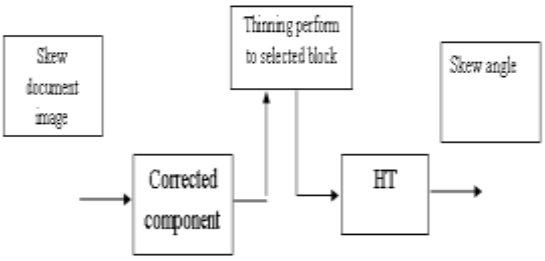
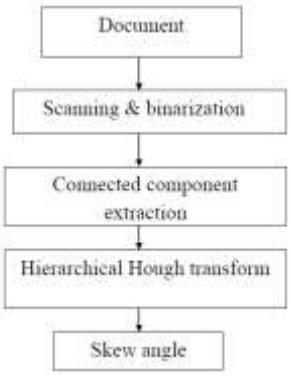


Fig. 4 skew detection and correction

Flow chart



This is a fast and efficient technique for skew detection and correction. This works very well in case of magazines, newspapers and in handwritten documents also. But the noise and variation in the document resolution are still the main problems in the devnagri skew detection and correction. There are various other methods for detecting and correcting the skew angle of the scan document with the different languages.

CONCLUSION

There are so many ways for detecting and correcting skew in a given document. Skew can be detected by various techniques like Fourier-transformation, nearest neighbor, cross- correlation, Hough transformation, moments, etc.

Every technique has some limitations, some of them provide us speed but are suitable only for small text, some provide us accurate results but are slow in speed, some are costlier if they are good in speed and accuracy. So we are going to propose a new technique with Gray Scale images, which will reduce the time and cost for detecting and correcting the skew in devnagri handwritten and printed document scanned images.

REFERENCES

[1] Adeline paiement, majid mirmehdi, senior member, ieee, xianghua xie, member, ieee, And mark c. K. Hamilton," integrated segmentation and interpolation Of sparse data" ieee transactions on image processing, vol. 23, no. 1, january 2014.

[2] Ruby singh1, Ramandeep kaur2, "Skew detection in image processing", Int.J.Computer Technology & Applications,Vol 4 (3),478-485 IJCTA | May-June 2013.

[3] A. Papandreou and B. Gatos," A Coarse to Fine Skew Estimation Technique for Handwritten Words" 2013 12th International Conference on Document Analysis and Recognition.

[4] Mamatha Hosalli Ramappa1 and Srikantamurthy Krishnamurthy, "Skew Detection, Correction and Segmentation of Handwritten Kannada Document" International Journal of Advanced Science and Technology Vol. 48, November, 2012.

[5] Dan WANG, Xichang WANG," A Skew Angle Detection Algorithm based on Maximum Gradient Difference" 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE) December 16-18.

[6] Shazia Akram, Dr. Mehraj-Ud-Din Dar, Aasia Quyoum, "Document Image Processing - A Review" International Journal of Computer Applications (0975 – 8887) Volume 10– No.5, November 2010

[7] Qi Xiaorui, Ma Lei, Sun Changjiang, Liu Jiang "Fast skew angle detection algorithm for scanned document images" vol.32,no.21,2006.11,pp.194-196