ISSN:2229-6093
Sunitha Parashar et al ,Int.J.Computer Technology & Applications,Vol 3 (4), 1619-1624

Finding skewness and deskewing scanned document

Sunita Parashar[1],Sharuti Sogi[2]
Department of Computer Sc. & Engineering, H.C.T.M, Kaithal, Haryana (India)
Sunita.tu@gmail.com[1],jindal.shruti88@gmail.com[2]

Abstract

Document image processing has become an increasingly important technology in the automation of office documentation tasks. Automatic document scanners such as text readers and OCR systems are an essential component of systems capable of those tasks. One of the problems in this field is that the document to be read is not always placed correctly on a flatbed scanner. This means that the document may be skewed on the scanner bed, resulting in a skewed image. Consequently, detecting the skew of a document image and correcting it are important issues in realizing a practical document reader.

1. Introduction

Organizations are moving at a fast pace from paper to electronic documents. However, large amounts of paper documents inherited from a recent past are still needed. Digitalization of documents appears as a bridge over the gap of past and present technologies. Scanners tend to be of widespread use for the digitalization of documents. One of the important problems in this field is that very often documents are not always correctly placed on the flat-bet scanner either manually by operators or by the automatic feeding device. This very frequent problem yields rotated images. For humans, rotated images are unpleasant for visualization and introduce extra difficulty in text reading. For machine processing, image skew brings a number of problems that range from needing extra space for storage to making more error prone the recognition and transcription of the image by automatic OCR tools. These reasons make skew detection and correction phases a common place in any environment for document processing. Very frequently the digitalization process of documents produce i images rotated of small angles in relation to the original image axis. The skew introduced makes more difficult the visualization of images by human users. Besides that, it increases the complexity of any sort of automatic image recognition, degrades the performance of OCR tools, increases the space needed for image storage, etc. Thus, skew correction is an important part of any document processing system being a matter of concern of researchers for almost two decades now. The search for faster and good quality solutions to this problem is still on.

2. Document Image Processing Steps

In document analysis the first step is to acquire a digitized raster image of the document using a suitable scanning system. Then it is followed by page layout analysis and character recognition. Before the structure of the text is obtained, a test is carried out to find out whether the document is skewed. Then skew is corrected and thereafter character recognition is done.

1. Scanning the Document
2. Skew Detection
3. **Skew Removal**
4.**Page Layout Analysis and Character Recognition**

Figure 1: Document Image Processing Steps

3. Document Skew and Its Reasons

The conversion of paper documents to electronic format is routinely done for record management, automated document delivery, document archiving, journal distribution etc. The stages of document conversion include scanning, displaying, image processing, text recognition, image and text database creation and quality assurance. During the scanning process, the whole document or a portion of it is fed through a loose-leaf page scanner. Some times pages are not fed properly into the scanner causing skew-ness of these bitmapped-image pages. A significant skew in document can be detected by human vision easily and the skew correction can be made by re-scanning the document, whereas for mild skew it may not be possible to notice its skew as human vision system fails to identify it. Even a smallest skew angle existing in a given document image results in the failure of segmentation of complete characters from words or a text lines, as the distance between the character reduces. Further most of the OCRs and document retrieval/display systems are very sensitive to skew in document images. Following figures shows skewed scanned document.

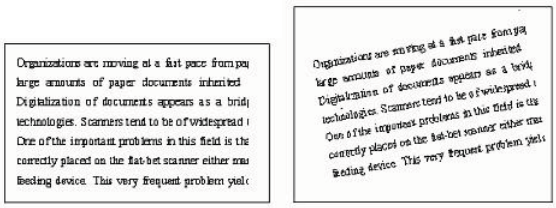

Figure 2: (a) original document and (b) skewed scanned Document

It is important to detect and correct skew ness. There can be many reasons for skewness in document images. But there are two most basic reasons are enumerated below. Skew in Scanning Process

1. **Skew in Scanning Process**
2. **Skewed HandWriting**
3. **Skewed Original Document**

Skewed Original Document: one of the basic reasons of skew ness can be that original source of scanning document may be skewed. This can happen when we are using type writer written document or print outs taken from printer where paper has not been placed properly.

3.1 Types of Document Skew

There are two type of skew in document images

1. **Single Skew**
2. **Multiple Skew**

Single Skew: In this skew, whole document is skewed to single angle. Most of document images have this type of skew-ness. This work deals with Single Skew problem.

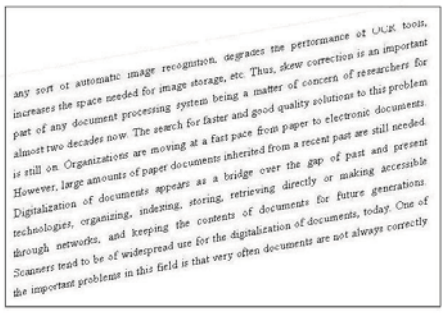

Figure 3: Single Skew

Multiple Skew: In this, scanned document can have many sections; each may be skewed to different angle. Detecting such type of skew-ness needs lot of efforts. Multiple Skew problem exists rarely and has not got lot attention from researchers.

Figure 4: Multiple Skew

4. Skew Detection Approaches

Several approaches have been proposed as alternatives for skew angle detection of document images. All of them require a dominant text area to be present in order to work properly. Main approaches for skew detection include:

- **Hough transform**
- **Projection Profile**
- **Nearest neighbor**
- **Principal Component Analysis**

Hough transform is a popular method for skew detection. It is capable of locating fragmented lines in a binary image. Therefore given a group of black pixels, one can find the imaginary line or lines that go through the maximum number of these pixels. Given a binary image with a dominant text area, the detected lines will most probably go along the whole middle zone of the textual lines. Hence these lines have approximately the same skew as the reference lines of the text which the skew of the whole page. Whenever the Hough transform is used, there is always a tradeoff between accuracy and speed. The more accurate the angles of the detected lines are, the more computation is required. In addition the computation time depends on the number of pixels in the image. Connected component analysis is required. Then several iterations take place in order to connect each component to its nearest neighbor in recursion. This process results in several chains representing the textual lines. A line that goes through the central mass of the components in a chain approximates the skew of the associated textual line. Alternatively, one can calculate the skew by averaging the skews of the lines that connect neighboring components in a chain. This method is based on the fact that inter character and inter word spaces between two consecutive characters or words respectively, are usually smaller than spaces between such neighboring elements that belong to different lines.

4.1 Hough Transform

The Hough transform is a technique which can be used to isolate features of a particular shape within an image. Because it requires that the desired features be specified in some parametric form, the classical Hough transform is most commonly used for the detection of regular curves such as lines, circles, ellipses, etc The **Hough transform** (pronounced h?if/, rhymes with tough) is a feature extraction technique used in image analysis, computer vision, and digital image processing. The purpose of the technique is to find imperfect instances of objects within a certain class of shapes by a voting procedure. This voting procedure is carried out in a parameter space, from which object candidates are obtained as local maxima in a so-called accumulator space that is explicitly constructed by the algorithm for computing the Hough transform. The classical Hough transform was concerned with the identification of lines in the image, but later the Hough transform has been extended to identifying positions of arbitrary shapes, most commonly circles or ellipses. The Hough transform as it is universally used today was invented by Richard Duda and Peter Hart in 1972, who called it a "generalized Hough transform". The simplest case of Hough transform is the linear transform for detecting straight lines. In the image space, the straight line can be described as y = mx + b and can be graphically plotted for each pair of image points (x,y). In the Hough transform, a main idea is to consider the characteristics of the straight line not as image points x or y, but in terms of its parameters, here the slope parameter m and the intercept parameter b. Based on that fact, the straight line y = mx + b can be represented as a point (b, m) in the parameter space. However, one faces the problem that vertical lines give rise to unbounded values of the parameters m and b. For computational reasons, it is therefore better to parameterize the lines in the Hough transform with two other parameters, commonly referred to as **?Ò**(rho) and **?Ò**(*theta*). The parameter **?•**represents the distance between the line and the origin, while **?•**is the angle of the vector from the origin to this closest point (see Coordinates)

Properties of Hough Transform

(1) Hough Transform is “Voting” algorithm.

(2) Since each point is handled independently, parallel implementations are possible.
(3) Hough is computationally expensive algorithm and so it is advised to preprocess the document to reduce the number of pixels in order to reduce the processing. So current research is going on to find the most appropriate pixels form document image to be used for determining orientation of document using Hough Transform

4.2 Projection Profile

The traditional projection profile approach was proposed by Postl . In this approach, the input document is rotated through a range of angles and a Projection profile is calculated at each angle Features are then extracted from each projection profile to determine the skew angle. As we know that baseline is the part of the script having most of the black pixels. Projection profile based technique takes the advantage of this property to find the space between the lines of the script. Then the tilt angle using projection and rotation can easily be estimated. Projection Profile based algorithm works as follows :
1. Start with (?û=0) and project horizontally and vertically to get the sum of the black pixels in the binary image and record the max of these two sums as well as ?Z.
2. Increase ?¹by 1 and rotate the image clockwise and project to get the maximum of the two sums. If the current sum is greater than the previous one then update the maximum sum and ?v other wise go to the next ?v.
3. Repeat Step 2 till ?Ö=90.
4. Rotate the image either by ?4or by ?4+90 depending on whether the maximum are gotten horizontally or vertically respectively.
In projection profiles, where a histogram is created at each possible angle and a 'cost function' is applied to this histogram. The skew angle is the angle at which this cost function is maximized.

4.3 Nearest Neighbor

Hashizume et al propose a method that computes the nearest neighbor of each character and creates a histogram of these values to detect the skew angle. . He found all the connected components in the document and computed the direction of its nearest neighbor for each component. A histogram of the direction angle is computed, the peak of which indicates the document skew angle. Liu et al use a similar technique to Hashizume et al, in which they detect and remove the characters with ascenders or descenders and then calculate the inclination between adjacent base points of connected components. These two methods rely on the information that in-line spacing is less than between-line spacing.

4.4 Principal Component Analysis

Principal component analysis (PCA) is a technique used to reduce multidimensional data sets to lower dimensions for analysis. Depending on the field of application, it is also named the **Hotelling transform** or **proper orthogonal decomposition (POD)**. PCA was invented in 1901 by Karl Pearson. Now it is mostly used as a tool in exploratory data analysis and for making predictive models. PCA involves the calculation of the eigen value decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for a given data in least square terms. PCA can be used for dimensionality reduction in a data set by retaining those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring higher order ones. Such low-order components often contain the "most important" aspects of the data. However, depending on the application this may not always be the case Principal Components Analysis (PCA) is a way of finding the directions along which a distribution exhibits the greatest variation . These directions are termed in the Principal Components of the distribution. They correspond to the most significant eigenvectors of the covariance matrix of the data points. PCA is used to find

the principal axis of the foreground distribution. This gives an angle for the baseline but not necessarily its vertical position. Steinherz used this method on Latin text. It was found that the performance is high when using the background pixels for PCA rather than the foreground. This method works as follows: First of all Convert the image into a set of vectors describing each foreground pixel that makes up the text. Then PCA is performed on these points and eigenvector with the largest eigen value is chosen. This vector gives the direction of the baseline. Image is rotated so that this estimated baseline angle lies horizontal. Projection histogram of the image is taken then and peak of this histogram is used to determine the vertical position of the baseline.

5. Applications of document image processing

5.1 Multimedia applications of Document Imaging

Document Imaging turns your computer into a virtual filing cabinet where any and all of your everyday forms, reports and other space-consuming papers can be stored conveniently and accessibly. This you may already know, but what you don't know is the extent to which imaging technology can be applied. The A&L Document Console extends its EMR capabilities considerably: the ability to not only store the contents of sheets of paper within your computer, but audio and video clips as well. Some of you may be familiar with common computer audio and video formats – the current publicity surrounding MP3, WMA, WAV, AAC, OGG , Real Media and various other formats being a prime example – but it is more than likely that even if an office wants to transfer its raw video or recorded audio to a computer-readable format, it lacks both the time and skill to see such an application through. Enter Document Console, which gives the user an easy-to-use interface through which sound and video storage can be done quickly and with a minimal knowledge of the technology behind such a process. Plus, with this new feature, it becomes possible to index every single item which enters your office in a single program, increasing overall organization and making sure you won't be misplacing important clips and such.

5.2 Document imaging for medical record

Our document imaging software takes a paper medical record and creates a universal electronic patient medical record that is easily managed. By scanning patient records into a searchable PDF file that is stored electronically on a network server, physicians and support staff can retrieve a patient's records from any networked computer. Our software also allows patient records to be retrieved easily over the Internet, from a hospital, a satellite office, or from home.

5.3 Banking and financial services

The Banking and Financial Service sectors provide significant opportunities for Document Imaging software and hardware including document scanners as this market is currently paper driven and prospects are looking for ways to improve their existing manual processes.

5.4 Real estate

Imaging and document management services can provide both commercial and residential real estate companies valuable advantages. Real estate transactions often result in hundreds of pages of documents. By capturing and storing these documents digitally, credit reports, mortgage forms, land plots, title information, and contracts can be easily and quickly located and displayed on any computer.

5.5 Transportation

Document imaging is a huge help in shipping/transportation industry. Weight bills, bills of lading, invoices, receiving documents can all be imaged, which facilitates faster movement of information among sites and faster retrieval of documents for reference.

6. References

[1]www.google.co.in
[2]Baird H. S., "The skew angle of printed documents", Proc. of SPSE 40th Symposium on Hybrid imaging systems, Rochester, NY, 1987, pp 739-743.
[3]Wikipedia Hough transform source URL:http://en.wikipedia.org/wiki/Hough_transform
[4]B. Yu and A. K. Jain, "A robust and fast skew detection algorithm for generic documents," Pattern Recognition, 29, no. 10, 1996, pp. 1599-1630.
[5]B. V. Dhandra, V. S. Malemath, Mallikarjun H, Ravindra Hegadi, "Skew Detection in Binary Image Documents Based on Image Dilation and Region labeling Approach", The 18th International Conference on Pattern Recognition (ICPR'06), 2006.
[6]Manjunath Aradhya V N, Hemantha Kumar G. and Shivakumara P, "Skew detection technique for binary document images based on Hough transform", International Journal of Information Technology, Vol. 3, 2006.
[7]Akiyama T and Hagita N, Automated entry system for printed documents, Pattern Recognition, Vol. 23, No. 11, 1990, pp 1141-1158.
[8]Baird H.S, The Skew Angle of Printed Documents, Proceedings of Conference Society of Photographic Scientists and Engineers, Rocherster, New York, 1987, pp 14-21.
[9]Cao Yang, Shuhua Wang, Li Heng., Skew detection and correction in document images based on straight-line fitting, Pattern Recognition Letters, 24, pp 1871-1879, 2003.
[10]Gonzales R.C and Woods R.E, Digital Image Processing, 2nd ed., Pearson Education Asia, 2002