Text Extraction from Document Images Based on Hough Transform Technique

Vikas K. Yeotikar, Manish T. Wanjarri, Jageshvar K. Keche, Dr. Mahendra P. Dhore

Abstract - Text extraction in document images has been developing rapidly since 1990s and is an important research field in content-based information indexing and retrieval, automatic annotation and structuring of document images. Extraction of text information involves detection, localization, tracking, extraction, enhancement, and recognition of the text from a given document images. However, variations of text due to differences in size, style, orientation, and alignment, as well as low image contrast and complex background make the problem of automatic text extraction extremely difficult and challenging job. A large number of techniques have been proposed to address this problem and the purpose of this paper is to classify and review Hough Transform techniques to extract text from document images. Hough Transform (HT) is recognized as a powerful tool for graphic element extraction from images due to its global vision and robustness in noisy or degraded environment. The method herein proposed detects text lines on document images which may include either lines oriented in several directions, erasures, or annotations between mainlines. At each stage of the process, the best text-line hypothesis is generated in the Hough Transform domain. Also we are using Edge Detection and Thresholding for text extraction from document images.

Key Words — Document Image Analysis (DIA), Feature Extraction Technique (FET), Hough Transform Technique(HTT), Edge Detection(ED) and Thresholding.

I. INTRODUCTION

Analysis of document images for information extraction has gained immense importance in recent past. Wide variety of information, which has been conventionally stored on paper, is now being converted into electronic form for better storage and intelligent processing. This needs processing of documents using image analysis algorithms. Locating text image blocks and tables, and defining appropriate algorithm is the major challenge in document image analysis [1][2]. In this paper we use Hough Transform, Edge Detection and Thresholding for text extraction from document images.

The Hough transform is a feature extraction technique used in image analysis, computer vision, and digital image processing.[3]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 ugh transform [4] after the related 1962 patent pf Paul Hough [5]. The transform was popularized in the computer vision community by Dana H. Ballard through a 1981 journal article titled "Generalizing the hough transform to detect arbitrary shapes.

Edge detection is a set of mathematical methods which aim at identifying points in digital document images at which the brightness of document image changes sharply or, more formally, has discontinuities. Edge detection is a fundamental tool in image processing, machine vision and computer vision, particularly in the areas of feature detection. Edge-detection techniques are used for performing document image segmentation tasks. This paper focuses on various edge detection techniques such as Sobel and Canny.

In document image processing, thresholding is a familiar technique for document image segmentation. The experimental results illustrate the importance and the usefulness of the approach for the specified class of document images. In this paper we have introduced general approach for document images by performing Global thresholding and Automatic thresholding.

II. HOUGH TRANSFORM TECHNIQUE

A number of methods have previously been proposed in the literature for identifying document image skew angles. Mainly, it is based on Hough transform. Most existing approaches use the Hough transform or enhanced versions Hough transform detects straight lines in an image. The algorithm transforms each of the edge pixels in an image space into a curve in a parametric space. The peak in the Hough space represents the dominant line and it’s skew. The major drawback of this method is that it is computationally expensive and is difficult to choose a peak in the Hough space when text becomes sparse.

A. Line selection in the Hough domain

The Hough transform is applied to the gravity centers of the connected components in the image. In the Hough domain, collinear alignments are searched in any direction. The process takes into account possible fluctuations of text lines, slight variations of the main direction, the irregularity of interlines and does not assume any privileged direction.

The Hough transform is a line to point transformation from the Cartesian space to the polar coordinate space. A line in the Cartesian coordinate space can be described by:
\[ p = x \cos \theta + y \sin \theta \]

where \( p \) is the normal distance of the line from the origin and \( \theta \) the angle between the x-axis and the normal line. A line corresponds to a point \((p, \theta)\) in the Hough domain which is quantized into cells. For each component gravity center in the image, the set of lines passing through that point for different discrete values of \( p \) and \( \theta \) corresponds to a set of cells in the
Hough domain. The cells are initialised to zero, and incremented by one, each time a point in the image (a gravity center) belongs to that line. Strong alignments correspond to cells with large values. [6]

### III. EDGE DETECTION TECHNIQUE

Edge detection techniques are generally used for finding discontinuities in gray level images. Edge detection is the most common approach for detecting meaningful discontinuities in the gray level. Document Image segmentation methods for detecting discontinuities are boundary-based methods. Edges are local changes in the document image intensity. Edges typically occur on the boundary between two regions. Important features can be extracted from the edges of an image (e.g., corners, lines, curves). Edge detection is an important feature for document image analysis. These features are used by higher-level computer vision algorithms (e.g., recognition). Edge detection is used for object detection which serves various applications like medical image processing, biometrics etc. Edge detection is an active area of research as it facilitates higher-level document image analysis. There are three different types of discontinuities in the grey level like point, line and edges. Spatial masks can be used to detect all the three types of discontinuities in a document image. [7] The most frequently used edge detection methods are used.

#### A. The Sobel Edge Detection

The Sobel operator is a well-known edge detector [8]. The original difference-based gradient computation is replaced by a Euclidean distance calculation. This vectorization of the algorithm allows for the effective use of the color information given that simple intensity differences would not represent differences between two color vectors as well as a Euclidean distance calculation.

The Sobel operator has been shown to be a good edge detector. In its expanded form, it will deal better with the information contained in color images without compromising it such as in methods where the operator is applied to each color plane independently [9]. In this case, the correlation between the various planes is lost and the final result would be probably less than adequate. The Sobel operator will suffer from an inability to identify all difference-based edges just as other Euclidean distance-based operators. The Sobel edge detection operator will be applied to the different color space.

#### B. The Canny Edge Detection

The Canny edge detector is regarded as one of the best edge detectors recently in use, Canny’s edge detector ensures good noise immunity and at the same time detects true edge points with minimum error. The Canny edge detector is an edge detection operator that uses a multi-stage algorithm to detect a wide range of edges in document images. It is developed by John Canny considered the mathematical problem of deriving an optimal smoothing filter given the criteria of detection, localization and minimizing multiple responses to a single edge. [10] He showed that the optimal filter given these assumptions is a sum of four exponential terms. He also showed that this filter can be well approximated by first-order derivatives of Gaussians.

Canny also introduced the notion of non-maximum suppression, which means that given the presmoothing filters, edge points are defined as points where the gradient magnitude assumes a local maximum in the gradient direction. Looking for the zero crossing of the 2nd derivative along the gradient direction was first proposed by Haralick, [11] It took less than two decades to find a modern geometric variational meaning for that operator that links it to the Marr-Hildreth (zero crossing of the Laplacian) edge detector. This observation was presented by Ron Kimmel and Alfred Bruckstein. [12]

Although his work was done in the early days of computer vision, the Canny edge detector (including its variations) is still a state-of-the-art edge detector. [13] Unless the preconditions are particularly suitable, it is hard to find an edge detector that performs significantly better than the Canny edge detector. The Canny-Deriche detector was derived from mathematical criteria as the Canny edge detector, although starting from a discrete viewpoint and then leading to a set of recursive filters for document image smoothing instead of exponential filters or Gaussian filters. [14]

### IV. THRESHOLDING TECHNIQUE

Many application of document image processing, the gray levels of pixels belonging to the object are quite different from the gray levels of pixels belonging to the background. A famous tool used in document image segmentation is thresholding. A variety of thresholding (also called as binarization) techniques includes both global and local thresholding. Thresholding becomes then a simple but effective tool to separate objects from the background. Thresholding is the simplest way to perform segmentation, and it is used in extensively in many image processing applications. Thresholding is based on the notion that regions correspond to the different regions can be classified by using range function applied to the intensity values of document image pixels. [15] Threshold segmentation techniques can be grouped into three different categories such as global, local and automatic thresholding.

#### A. Global Threshold

The simplest of all thresholding techniques is to partition the document image histogram by using a single global threshold. Segmentation is then accomplished by scanning the document image pixel by pixel and labeling each pixel as object or background, depending on whether the gray level of that pixel is greater or less than the value. The success of this method depends entirely on how well the histogram can be partitioned.

In general of all thresholding techniques is to partition the document image histogram by using a single global threshold T. Segmentation is accomplished by scanning the document image pixel by pixel and labeling each pixel as object or background, depending on whether the gray level of that pixel is greater or less than the value of T. When T depends only on f(x, y) (i.e. only on gray level values) the threshold is called global threshold.

For choosing a threshold automatically, Gonzalez and Woods describe some of the following procedures

1. Select an initial estimate for T.
ii) Segment the image using T. This will produce two groups of pixels: G1, consisting of all pixels with intensity values > T, and G2, consisting of pixels with values < T.

iii) Compute the average intensity values μ1 and μ2 for the pixels in the regions G1 and G2.

iv) Compute a new threshold value

\[ T = \frac{1}{2} (\mu_1 + \mu_2) \]

v) Repeat steps 2 through 4 until the difference in T in successive iterations is smaller than a predefined parameter T₀.

B. Automatic Threshold

Thresholding based document image segmentation requires finding a threshold value T that establishes the ‘border’ between graylevel document image range corresponding to objects and a range equivalent to background. After thresholding the graylevel document image is converted to binary. There exist algorithm that use more than one threshold values (multithresholding), which enables to assign pixels to one of a few classes instead of just two. Threshold value(s) may be entered manually or automatically.

Automatically selected threshold value for each document image by the system without human intervention is called an automatic threshold scheme. This is requirement the knowledge about the intensity characteristics of the objects, size of the objects, fractions of the document image occupied by the objects and the no of different types of objects appearing in the document image. [16, 17].

V. EXPERIMENTAL RESULTS

Historically, certain computer programs were written using only two digits rather than four to define the applicable year. Accordingly, the company’s software may recognize a date using “00” as 1900 rather than the year 2000.

a) Original Document Image   b) Canny edge detection

Historically, certain computer programs were written using only two digits rather than four to define the applicable year. Accordingly, the company’s software may recognize a date using “00” as 1900 rather than the year 2000.

Historically, certain computer programs were written using only two digits rather than four to define the applicable year. Accordingly, the company’s software may recognize a date using “00” as 1900 rather than the year 2000.

c) Sobel edge detection     d) Houghlines

Fig. 1. Hough Transform

Fig. 2. Hough Transform Function

Fig. 3. Line Detection Using Hough Transform

Historically, certain computer programs were written using only two digits rather than four to define the applicable year. Accordingly, the company’s software may recognize a date using “00” as 1900 rather than the year 2000.

a) Original Document Image   b) Canny edge detection

Historically, certain computer programs were written using only two digits rather than four to define the applicable year. Accordingly, the company’s software may recognize a date using “00” as 1900 rather than the year 2000.

c) Sobel edge detection     d) Houghlines

Fig. 1. Hough Transform

Fig. 2. Hough Transform Function

Fig. 3. Line Detection Using Hough Transform

Historically, certain computer programs were written using only two digits rather than four to define the applicable year. Accordingly, the company’s software may recognize a date using “00” as 1900 rather than the year 2000.

a) Original Document Image   b) Canny edge detection

Historically, certain computer programs were written using only two digits rather than four to define the applicable year. Accordingly, the company’s software may recognize a date using “00” as 1900 rather than the year 2000.

c) Sobel edge detection     d) Houghlines

Fig. 1. Hough Transform

Fig. 2. Hough Transform Function

Fig. 3. Line Detection Using Hough Transform

Historically, certain computer programs were written using only two digits rather than four to define the applicable year. Accordingly, the company’s software may recognize a date using “00” as 1900 rather than the year 2000.
Fig. 4. Search for line segments in an image and highlight the longest segment.

a) Original Document Image

b) Segmented result using global thresholding

c) Original Document Image

d) Automated Thresholding

Figure 1 shows a) Original Document Image, b) Normal axis & c) Hough Matrix, simple way to performing Hough transform.

Figure 2. Shows a) Original Document Image, b) Canny edge detection, edges by looking for local maxima of the gradient of $f(x, y)$. The gradient is calculated using the derivative of a Gaussian filter. The method uses two thresholds to detect strong and weak edges, and includes the weak edges in the output only if they are connected to strong edges. Therefore, this method is more likely to detect true weak edges. c) Sobel edge detection, edges using the Sobel approximation to the derivatives. & d) Hough line, using the Hough line to find the location of all nonzero pixels in the document image that contributed to that peak and construct line segments based on those pixels.

Figure 3. shows a) Line detection, using the function of Hough on a simple binary image. b) Hough detection & c) Hough Transform, The MATLAB has a function called Hough that computes the Hough Transform and display the new curve on the Hough transform.

Figure 4. Shows a) Original Document Image, b) Hough Longest Segment, determine the end points of the longest line segments & c) Highlights the longest line segment.

Figure 5. a) Shows the original document image, b) shows the segmented result by using global thresholding (the border was added manually for clarity) & c) shows the automatic thresholding i.e. for without human intervention called an automatic threshold scheme.

**CONCLUSION**

A text-image-analysis is needed to enable a text information extraction system to be used for any type of document image.

We have successfully implemented the Hough Transform, Edge Detection and Thresholding technique. In edge detection we performed on sobel and canny method. Thresholding is implemented using Global and Automatic method. We got results on document images by implementing above technique in MATLAB.

**REFERENCES**


