Volume 4, Issue 11, November 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Pre-processing Techniques in Character Recognition

Purna Vithlani
Department of Computer Science Saurashtra University,
Rajkot, Gujarat, India

Abstract— Pre-processing is the basic phase of character recognition. This paper deals with the various pre-processing techniques like thresholding, noise removal, skew detection and correction.

Keywords— Character Recognition, Thresholding, Noise Removal, Skew Detection and Correction

I. Introduction

Pre-processing is necessary to modify the raw data to correct deficiencies in the data acquisition process due to limitations of the capturing device sensor. Data pre-processing describes any type of processing performed on raw data to prepare it for another processing procedure. Pre-processing is the preliminary step which transforms the data into a format that will be more easily and effectively processed. Thus, pre-processing is an essential stage since it controls the suitability of the results for the successive stages. The stages in character recognition are in a pipeline fashion meaning that each stage depends on the success of the previous stage in order to produce optimal results.

The main objective of the pre-processing stage is to normalize and remove variations that would otherwise complicate the classification and reduce the recognition rate.

II. FACTORS AFFECTING ACCURACY OF CHARACTER RECOGNITION

Following factors affect the accuracy of character recognition.

- Older or discolored documents
- Low contrast documents
- Scanner quality
- Type of printed document
- Paper quality
- Fonts used in the document
- Scan Resolution - The recommended best scanning resolution for OCR accuracy is 300 dpi. Higher resolutions do not necessarily result in better accuracy and can slow down OCR processing time. Resolutions below 300 dpi may affect the quality and accuracy of OCR results.

III. PRE-PROCESSING TECHNIQUES

A series of operations have to be performed during the pre-processing stages. The main objective of the pre-processing is to organize the information so that the subsequent character recognition task becomes simpler. Pre-processing includes following stages:

A. Thresholding

Thresholding converts gray scale image to binary image (black and white image). The goal of Thresholding is to remove only the background, by setting it to white, and leave the foreground image unchanged. Thresholding selects a proper threshold value for the image and then convert all the pixels above the threshold to white and below the threshold to black.

In any image analysis or enhancement problem, it is very essential to identify the objects of interest from the rest. Thresholding is the process of separating the objects of an image from its background.

Otsu’s Thresholding method is an efficient and frequently used method. Otsu's method is an image processing technique that can be used to convert a gray scale image into a binary image by calculating a threshold to split pixels into two classes. More generally, Otsu's method can be used to split a histogram into two classes which minimizes the intra-class variance of the data contained within the class. The technique is named after Nobuyuki Otsu who published the technique in the 1979 paper "A Threshold Selection Method from Gray-Level Histograms".

Otsu’s method uses the image histogram data as input and finds a pixel value (threshold level) that is able to separate the image into foreground and background (or even to multiple levels). The algorithm assumes that the image contains two classes of pixels following bi-modal histogram (foreground pixels and background pixels), it then calculates the optimum threshold separating the two classes so that their combined spread (intra-class variance) is minimal. It converts gray scale image into a binary image on the basis of pixel whether it is below or above the specified threshold value.

B. Noise Removal

Scanned documents often contain noise that arises due to printer, scanner, print quality, age of the document, etc. Therefore, it is necessary to filter this noise before character recognition.

Image noise is random variation of brightness or colour information in images, and is usually an aspect of electronic noise. It can be produced by the sensor and circuitry of a scanner or digital camera. Image noise is an undesirable by-product of image capture that adds spurious and extraneous information. Noise is visible as a grain or film grain in an image. Optical scanning devices introduce some noises like, disconnected line segments, bumps and gaps in lines, filled loops etc. It is necessary to remove all these noise elements prior to the character recognition.

Noise removal is the process of removing or reducing unwanted noise.

Types of Noises

Depending on the type of disturbance, the noise can affect the image to different extent. There are following types of noises.

1) *Gaussian Noise (Amplifier Noise):* Gaussian noise is caused by random fluctuations in the signal. It is modelled by random values added to an image. In Gaussian noise, each pixel in the image will be changed from its original value by a small amount. Each pixel in the noisy image is the sum of the true pixel value and a random, Gaussian distributed noise value.
2) *Salt and Pepper Noise:* Salt and pepper noise is also called fat-tail distributed or impulsive noise or spike noise. An image containing salt-and-pepper noise will have dark pixels in bright regions and bright pixels in dark regions. It presents itself as sparsely occurring white and black pixels. This noise arises in the image because of sharp and sudden changes of image signal. An effective noise reduction method for this type of noise is a median filter or a morphological filter.
3) *Shot Noise:* Shot noise is the noise that can cause, when number of photons sensed by the sensor is not sufficient to provide detectable statistical information. This noise is known as photon shot noise. Shot noise has a root-mean-square value proportional to the square root of the image intensity, and the noises at different pixels are independent of one another. Shot noise follows a Poisson distribution, which is usually not very different from Gaussian.

Noise Removal Methods

1) *Median Filter:* Median Filter is a non-linear method. The median filter is effective for removing salt and pepper noise. The main idea of the median filter is to run through the signal entry by entry, replacing each entry with the median of neighbouring entries. The pattern of neighbours is called the "window", which slides, entry by entry, over the entire signal. Note that if the window has an odd number of entries, then the median is simple to define: it is just the middle value after all the entries in the window are sorted numerically. If the window contains an even number of pixels, the average of the two middle pixel values is used.

   The median filter takes an area of an image (3x3, 5x5, 7x7, etc.), sorts out all the pixel values in that area, and replaces the centre pixel with the median value. Figure illustrates an example of how the median filter is calculated.

123	127	150	120	100
119	115	134	121	120
111	120	122	125	180
111	119	145	100	200
110	120	120	130	150

		121		

The sorted pixel values of the shaded area are: (100,115,119,120,121,122,125,134,145). So that median value is 121.

2) *Mean Filter:* Mean filtering is a simple and easy to implement method of reducing noise from an image. Mean filter is an averaging linear filter. The idea of mean filtering is simply to replace each pixel value in an image with the mean 'average' value of its neighbors, including itself. This has the effect of eliminating pixel values which are unrepresentative of their surroundings. Mean filtering is usually thought of as a convolution filter. Like other

convolutions it is based around a kernel, which represents the shape and size of the neighborhood to be sampled when calculating mean. Often 3*3 square kernel/mask is used. Mean filters show very good performance for the removal of many noise types (e.g. Gaussian noise).

123	127	150	120	100
119	115	134	123	120
111	121	122	125	180
111	120	155	101	200
110	120	120	130	150

		124		

Mean value of (115,134,123,121,122,125,120,155,101) is 124. So that mean value is 124.

C. Skew Detection and Correction

Skew detection and correction of scanned document images is one of the most important stages of pre processing. The skew of the scanned document image specifies the deviation of its text lines from horizontal or vertical axis. The skew of the document image can be a global (all document's blocks have the same orientation), multiple (document's blocks have a different orientation) or non uniform (multiple orientation in a text line).

Skew correction aligns an image before processing because text segmentation and recognition methods require properly aligned text lines.

Many methods for skew detection and correction of scanned document images have been proposed. These methods include projection profile analysis, Hough transform, Scan line, nearest neighbour clustering etc.

Skew Detection and Correction Methods

1) *Projection Profile Analysis:* In this method, the horizontal or vertical projection profile is used as a suitable feature for skew detection. Horizontal (or vertical) projection profile is the histogram of a one-dimensional array with a number of entries equal to the number of rows (or columns). The number of black pixels in a row (or column) is stored in the corresponding entry. The horizontal projection profile is based on the histogram of black pixels along horizontal scan lines. For a script with horizontal text lines, the horizontal projection profile will have peaks at text line positions and troughs at positions in between successive text lines.

The projection profile analysis process is as follows:

1. Dimension reduction: rotate the binary input image to different angles and at any angle do "a" and "b".
   a. Feature extraction: obtain the projection profile.
   b. Feature extraction: calculate criterion function.
2. Skew estimation: obtain the angle corresponding to the maximum value of criterion function.

2) *Hough Transform:* The Hough transform is used to detect lines, circles or other parametric curves. It was introduced in 1962 (Hough 1962) and first used to find lines in images a decade later (Duda 1972). The goal is to find the location of lines in images. The straight lines corresponding to the image text lines are extracted as features for skew detection. Deviation of image text lines from the horizontal or vertical axis is specified as its skew.

Hough Transform approach is preferred when the objective is to find lines or curves formed by groups of individual points on an image plane. The method involves a transformation from an image plane to a parameter space. Consider the case in which lines are the objects of interest. The line is expressed as

$$\rho = X \cos\theta + Y \sin\theta$$

There are two line parameters namely, the distance (ρ) and the angle (θ) which defines transformation space. Each coordinate (x, y) of ON pixel in the image plane is mapped onto the locations in the transformed plane for all possible straight lines. For all possible values of ρ and θ the transformations intersect at the same point on the transformed plane when multiple points are collinear. Therefore, the point (ρ, θ), which has the greatest accumulation of mapped points, indicates lines with these parameters. In practice, due to discretization error and noise, points mapped will not be exactly collinear. Thus the points do not map on to exactly the same location on the transformed plane. For connected lines or positions of lines, computations can be reduced greatly by considering not all (ρ, θ) points but only those (ρ, θ) points that are in one orientation as indicated by the angle.

Hough transform is performed on the scanned document image and the variance in P values is calculated for each value of θ. The angle that gives the maximum variance is the skew angle.

Hough Transform based method has the following step:

1. Dimension reduction: do "a" and "b" for the binary input image.
   a. Feature extraction: using the Hough Transform to find the text lines of the image.
   b. Feature extraction: calculate the criterion function for the angle θ.
2. Skew estimation: obtain the angle corresponding to the maximum value of criterion function.

Hough Transform is simple and easy to implement. It handles missing and occluded data very gracefully. It can be adapted to many types of forms, not just lines.

3) *Scan Line:* In this method the image is projected at several angles and the variance in the number of black pixels per projected scan line is determined. The angle at which the maximum variance occurs is the angle of skew.

Algorithm:

a. Calculate the coordinates, in the image plane, for each of the parallel scan lines that lie at a slope tan (θ) in the image plane. The coordinates are calculated using the Bresenham’s Line Drawing Algorithm.
b. For each scan line, count the number of non-background pixels that lie on the line.
c. Calculate the variance v in the number of black pixels that lie on each scan line for a given angle θ.
d. The angle of skew θ is given by the angle at which the maximum variance v max is found.

4) *Nearest Neighbour Clustering:* In this method, vectors connecting the image connected components to the nearest neighbours are used as features for skew detection. Nearest neighbours of each connected component are usually adjacent letters on the same text line. Therefore, the vectors connecting any connected component to its nearest neighbours are usually characterized by a line parallel to a text line. Since the deviation of the text lines from the horizontal or vertical axis defines the skew, text lines and parallel lines to them can be used as a convenience feature for skew detection. Therefore, those vectors can be used as features for skew detection.

The Nearest neighbour clustering process is as follows:

1. Dimension reduction: do "a", "b" and "c" for the input binary image.
   a. Feature selection: obtain the connected components.
   b. Feature extraction: determine the nearest neighbour set of each connected component. Obtain angle of vectors connecting each component to its nearest neighbour.
   c. Feature extraction: calculate the value of the criterion function.
2. Skew estimation: obtain the angle corresponding to the maximum value of criterion function.

IV. CONCLUSION

In this paper Pre-processing techniques used in document images as an initial step in character recognition systems were presented. Thresholding, types of noises, noise removal methods and skew detection and correction methods are discussed. The pre-processing techniques discussed so far give considerably good results.

REFERENCES

[1] http://cdn.intechopen.com/pdfs-wm/11405.pdf
[2] http://en.wikipedia.org/wiki/Binary_image
[3] http://en.wikipedia.org/wiki/Otsu's_method
[4] http://en.wikipedia.org/wiki/Image_noise
[5] http://en.wikipedia.org/wiki/Gaussian_noise
[6] http://en.wikipedia.org/wiki/Median_filter
[7] http://www.scientificbulletin.upb.ro/rev_docs_arhiva/full4365.pdf