Chapter: 3

Character Recognition System

Introduction

3.1 Steps of Bangla Character Recognition

3.2 Input Discrepancy

3.3 Optical Scanning

3.4 Location Segmentation

3.5 Preprocessing

3.5.1 Binarization

3.5.1.1 Otsu’s Method

3.5.1.2 Niblack’s Method

3.5.1.3 Sauvola Method

3.5.1.4 Bernsen

3.5.1.5 Local Maxima and Minima

3.5.1.6 Adaptive Contrast

3.5.1.7 Global-to-Local Approach

Introduction

Optical character recognition (OCR) is process of classification of optical patterns contained in a digital image corresponding to alphanumeric or other characters. The character recognition is achieved through important steps of segmentation, feature extraction and classification. OCR has gained increasing attention in both academic research and in industry. In this chapter we have collected together the basic ideas of OCR needed for a better understanding of this thesis.

Optical character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. A typical OCR system contains three logical components: an image scanner, OCR software and hardware and an output interface. The image scanner optically captures text images to he recognized. Text images are processed with OCR software and hardware. The process involves three operations: document analysis (extracting individual character images), recognizing these images (based on shape) and contextual processing (either to correct misclassifications made by the recognition algorithm or to limit recognition choices). OCR software attempts to identify characters by comparing shapes to those stored in the software library or database. The software tries to identify words using character proximity and will try to reconstruct the original page layout. High accuracy can be obtained by using sharp, clear scans of high-quality originals. The output interface is responsible for communication of OCR system results to the outside world.

3.1 Steps of Bangla Character Recognition Process

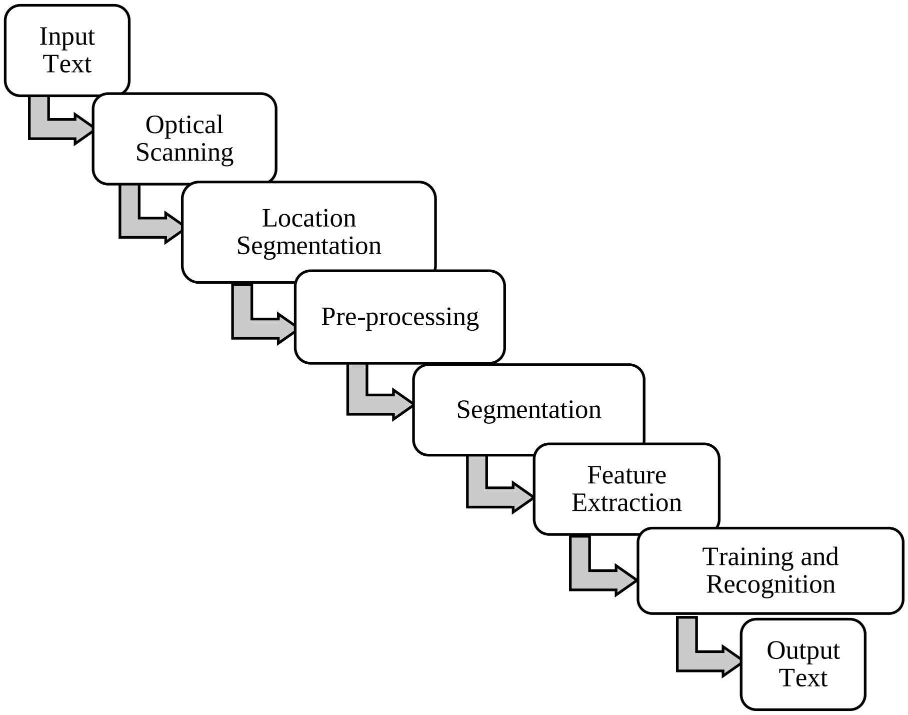


Fig. 3.1: The Components of OCR system

3.2 Input Discrepancy

Different types of inputs are feed into the OCR system. The types are shown in Fig. 3.2, we have used handwritten character in this thesis.

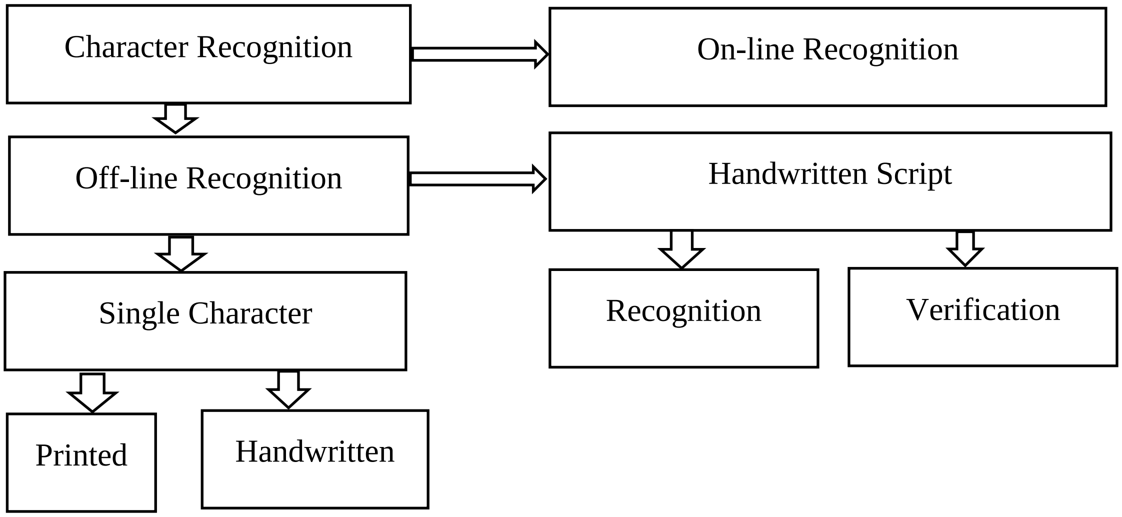


Fig.3.2: The different areas of character recognition

3.3 Optical Scanning

The first component in OCR is optical scanning. Through scanning process digital image of original document is captured. In OCR optical scanners are used which consist of transport mechanism and sensing device that converts light intensity into grey levels. Printed documents consist of black print on white background. When performing OCR multilevel image is converted into bi-level black and white image. This process known as thresholding is performed on scanner to save memory space and computational effort. The thresholding process is important as the results of recognition are totally dependent on quality of bi-level image. A fixed threshold is used where gray levels below this threshold are black and levels above are white. For high contrast document with uniform background a pre-chosen fixed threshold can be sufficient. However, documents encountered in practice have rather large range. In these cases more sophisticated methods for thresholding are required to obtain good results. The best thresholding methods vary threshold adapting to local properties of document such as contrast and brightness. However, such methods usually depend on multilevel scanning of document which requires more memory and computational capacity.

3.4 Location Segmentation

The next OCR component is location segmentation. Segmentation determines constituents of an image. It is necessary to locate regions of document which have printed data and are distinguished from figures and graphics. For example, when performing automatic mail sorting through envelopes address must be located and separated from other prints like stamps and company logos, prior to recognition. When applied to text, segmentation is isolation of characters or words. Most of OCR algorithms segment words into isolated characters which are recognized individually. Usually segmentation is performed by isolating each connected component. This technique is easy to implement but problems arise if characters touch or they are fragmented and consist of several parts. The main problems in segmentation are: (a) extraction of touching and fragmented characters (b) distinguishing noise from text (c) misinterpreting graphics and geometry with text and vice versa.

3.5 Pre-processing

The raw data depending on the data acquisition type is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Preprocessing are the steps performed before feature extraction, It consists of Binarization of the image, removal of noise, skew detection, segmentation, scaling. These steps are given in Fig. 3.3,

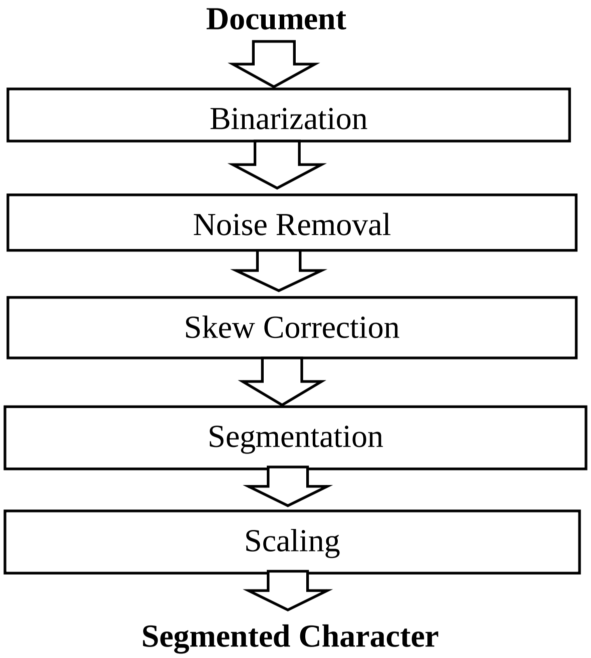


Fig.3.3: Pre-processing steps of image

3.5.1 Binarization

Binarization is the process of converting a pixel image to a binary image .Binarization of scanned images is the first step in most document image analysis systems. Selection of an appropriate binarization method for an input image domain is adifficult problem. Different Binarization algorithm gives different performances on different data sets. This is especially true in the case of historical documents with variation in contrast and illumination, smearing and smudging of text. The quality of the image has a significant impact on the OCR Performance. Noise is detected and eliminated during Binarization. So, selection of appropriate binarization algorithm is very important for OCR performance. Some binarization methods are-Otsu’s Method, Niblack’s Method, Sauvola Method, Bernsen, Local Maxima and Minima, Adaptive Contrast, Global-to-Local Approach, Morphological Contrast Intensification.

3.5.1.1 Otsu's Method

Otsu method is a global thresholding method that converts gray scale image into bi-level image. This technique divides the pixels into two classes one is foreground and the other is background. It chooses an optimal threshold that separates the image into two different classes. The threshold value is chosen such that the within-class variance is minimized and the between-class variance is maximized. The weighted within class variance of two classes is given as

$$\sigma^2 P(t) = P1(t)\sigma_1^2 + P2(t)\sigma_2^2 \tag{3.1}$$

Otsu method gives its best performance for only those images that have clear bi-modal pattern. But, degraded documents normally don't have such clear-cut pattern. Besides this, it does not perform well for images with uneven illumination and shadow.

3.5.1.2 Niblack's Method

Niblack [11] is a local thresholding method. In local thresholding methods, a different threshold value is calculated for each and every pixel. It uses local statistics of the image, such as variance, range to calculate the threshold. In Niblack method a rectangular window is slided over the gray scale image to estimate threshold of the pixels. It uses the local statistics mean and standard deviation of the window to estimate the threshold. Threshold $T(i,j)$ is estimated as

$$T(i,j) = \mu + k \times \sigma \tag{3.2}$$

Here, μ represents the mean of the window and σ represents the standard deviation of the window. The value of k is a constant and it defines the size and quality of binarization. As this method is dependent upon the local features of the image, it gets affected by blank areas in the image and is also not efficient for the images with background noise.

3.5.1.3 Sauvola Method

Sauvola method [11] is the improvement of the Niblack method. It is local variance method that uses standard deviation. Threshold is calculated as

$$T(i,j) = \mu \times [\,1 + k\frac{\sigma}{R} - 1] \tag{3.3}$$

Here, μ is the mean and σ is the standard deviation of the window. Values suggested for k and Rare 0.5 and 128. The window size and value of k will affect the quality of image but R will have very little affect. This method is used for documents having uneven illumination, light texture and stained documents. But, Sauvola method thins the text after its application.

3.5.1.4 Bernsen Method

Bernsen method [11] uses the contrast of the image. The threshold is estimated as the average of highest and lowest intensity values in the window. The Eq. 4 calculates the local contrast of the window

$$C(i.j) = I_{max} - I_{min} \tag{3.4}$$

The pixels are categorized as foreground or background by comparing the local contrast with threshold value. The pixel will be classified as background, if the local contrast is found to be less than the threshold and vice-versa. Bernsen method doesn't perform well for the images having more complex background.

3.5.1.5 Local Maxima and Minima

This method [11] uses contrast which depends upon the local minimum and maximum. A normalization factor is introduced which will compensate the effect of variation in image background. Image contrast is calculated as

$$C(i.j) = \frac{I_{max} - I_{min}}{I_{max} - I_{min} + \epsilon} \tag{3.5}$$

High contrast image pixels are found from the contrast image. Then local thresholding is performed with threshold value calculated from the found high contrast image pixels. It is not suitable for the bright text with proper bright background.

3.5.1.6 Adaptive Contrast

Adaptive contrast [12] is the consolidation of image contrast and image gradient. Adaptive contrast method removes the over-normalization problem of Local maxima minima method. Contrast map is constructed and binarized. Canny edge map is also constructed and combined with the resultant of the first step. This detects the actual text stroke edges. The text is extracted using local thresholding. It is estimated from the mean and standard deviation of the detected text stroke pixels within a window. The limitation of this particular method is that the canny edge detector extracts the fake edges also.

3.5.1.7 Global-to-Local Approach

Global thresholding followed by local thresholding is performed to binarize the image. In pre-processing Gaussian filter is used. Edge map is constructed using canny edge detector. Histogram of the pixels other than the edge pixels is obtained. The lower of the two clear peaks of the histogram is chosen as global threshold and the pixel values exceeding this threshold are turned background. Remaining pixels are classified as foreground or background depending upon the local threshold. Local threshold is estimated using the average of the highest and lowest gray value of the window. Shrink and Swell filters are used to amend the final results of binarization in post-processing [13].

3.5.1.8 Morphological Contrast Intensification

Gray scale morphological tools are used for background estimation and contrast intensification. It is a hybrid approach of global and local thresholding. Initially size of structuring element is chosen. The histogram of the distance between the consecutive edges is used to get the size of the structuring element. Background of the image is estimated using morphology. Global threshold is estimated by morphological contrast intensification. Text regions are separated with the help of histogram of contrast image. Local thresholding is performed to binarize the document image. Post processing is done for removing the noise from the binarized image.

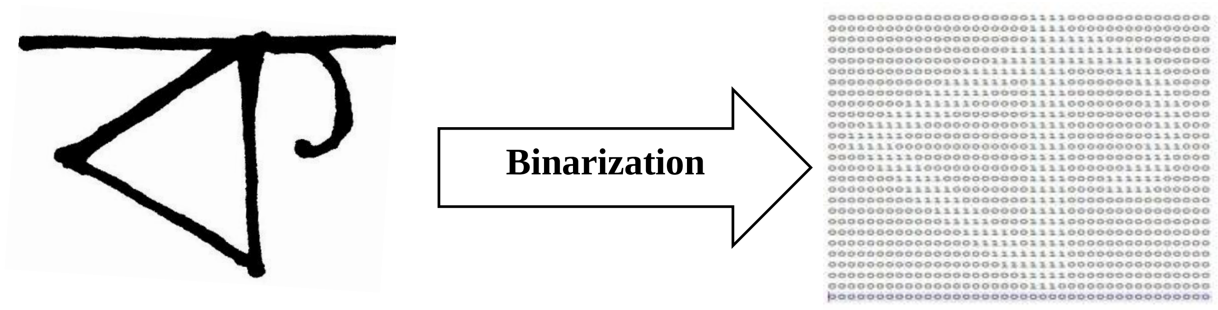


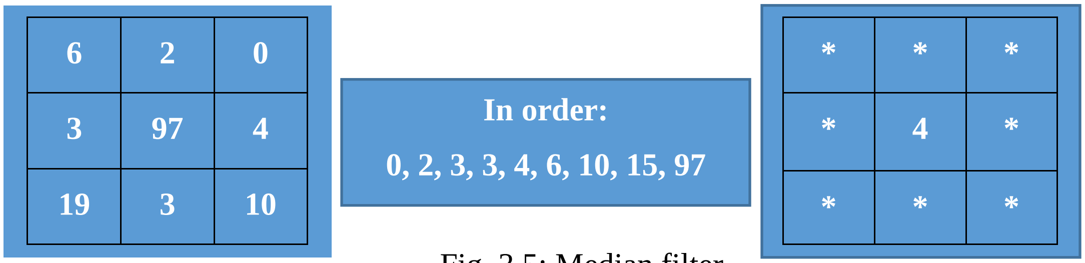
Fig. 3.4: Binalization process

3.5.2 Noise Removal

Scanned images may contain noise that emerges as a result of the printer, scanner, print quality, age of the document, etc. Therefore, it is imperative to filter this noise before dealing with the image. Low-pass filter is generally used to smoothen the characters for later processing.

3.5.2.1 Median Filter

The Median Filter is a nonlinear filtering technique, often used to remove noise. Such noise reduction is a typical pre-processing step to improve the results of later processing. This filtering Procedure is used to examine a sample of the input. It is performed using a window consisting of an odd number of input data samples. Median filtering is very widely used in digital image processing, particularly useful to reduce noise, salt and pepper noise. It is one kind of smoothing technique as well. All smoothing, techniques are effective at removing noise in smooth patches or smooth regions of a signal, but adversely affect edges.



6	2	0
3	97	4
19	3	10

In order:

0, 2, 3, 3, 4, 6, 10, 15, 97

*	*	*
*	4	*
*	*	*

Fig. 3.5: Median filter.

3.5.2.2 Gaussian Filter

A Gaussian filter is a filter whose impulse response is a Gaussian function. Gaussian filters are designed to give no overshoot to a step function input while minimizing the rise and fall time. This behavior is closely connected to the fact that the Gaussian filter has the minimum possible group delay. Gaussian filtering is a linear convolution algorithm unrelated to Median filter. The one-dimensional Gaussian filter has an impulse response given by,

$$g(x) = \sqrt{\frac{a}{\pi}} * e^{-ax^2} \tag{3.6}$$

Or, with the standard deviation as parameter: $\pi\sigma$

$$g(x) = \frac{1}{\sqrt{2\pi\sigma}} * e^{\frac{x^2}{2\sigma^2}} \tag{3.7}$$

In two dimensions, it is the product of two such Gaussians, one per direction:

$$g(x) = \frac{1}{2\pi\sigma^2} * e^{\frac{x^2+y^2}{2\sigma^2}} \tag{3.8}$$

Where, x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis, and σ is the standard deviation of the Gaussian distribution. Gaussian filtering allow user to make fine adjustment to the amount of partial averaging that occurs in the image.

3.5.3 Skew Detection and Correction

Skew is basically an angle that is created due to an angular Placement of document in the scanner. It can be corrected in two steps-

1. Estimation of skew angle θ_s, and
2. Rotation of image by θ_s, in the opposite direction.

Many skew detection and correction algorithms are available. At present, skew detection methods can he roughly classified as follows:

- The method based on Hough transforms.
- The skew detection method based on the analysis of the texture complexity.
- The method based on Cross Correlation.
- The method based on the Projection profile.
- The method based on Fourier transformation.
- The K nearest neighbor (K-NN) cluster method.

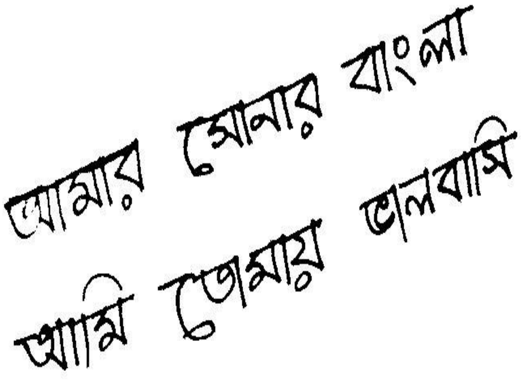


Fig. 3.6: Skew detection

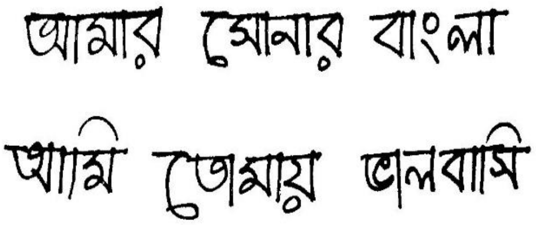


Fig. 3.7 Skew correction

3.5.4 Segmentation

Segmentation is a mechanism that disintegrates an image of sequence of characters into separate characters. This is the most vital and important portion for designing an efficient Bangla OCR because fie extraction and recognition process depends on this phase to make the recognition process successful. The output of this phase 51s consists of individual images of basic, modified and pound characters. Segmentation process includes the following steps. They are-

- Line Segmentation
- Word Segmentation
- Character Segmentation

3.5.4.1 Line Segmentation

Text line segmentation has been performed by scanning the input image horizontally and by keeping record of the number of black pixels in each row. Upper boundary of a line is the first row where the first black pixel is found. After finding the upper boundary, it continues scanning until a row whose next two consecutive rows have no black pixels, which is the lower boundary of the text line. It is noted that there exist more than two blank rows between two lines. The line detection process is shown in the Figure-13. And the various boundaries of the text lines are shown in Fig. 3.8.

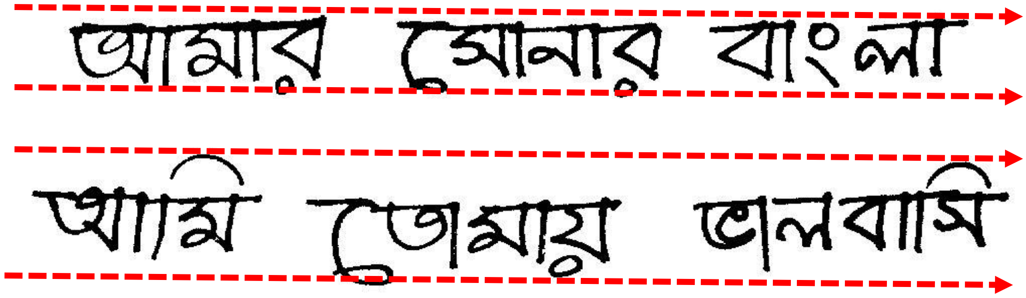


Fig. 3.8: Line Segmentation

3.5.4.2 Word Segmentation

After detecting a line, the system scans the image vertically from the upper boundary line to the lower boundary line of a text line. The number of black pixels in each column is counted. Starting boundary of a word is the first column where the first black pixel is found. After finding the starting boundary, it continues scanning until a column whose next two consecutive columns have no black pixels, which is the ending boundary of the word being processed. It is noted that there exist more than two blank columns between two words. Fig. 3.9 shows the word

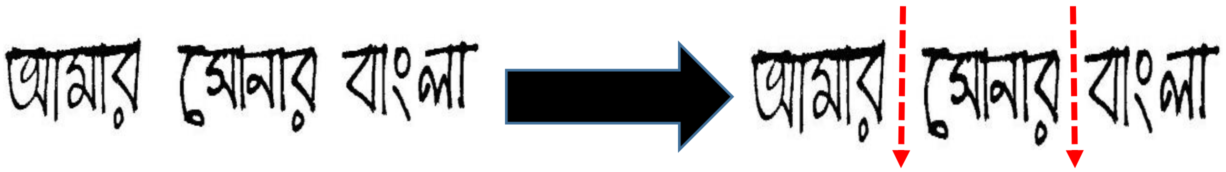


Fig. 3.9: Word Segmentation

3.5.4.3 Character Segmentation

To segment the individual characters in a word a vertical scan is performed from the upper boundary line of a word to the lower boundary line. If we reach the lower boundary line without facing any black pixel during scan then this column is assumed that the starting/ending boundary line of a character in the word. Vertical scanning is applicable if two consecutive characters are not connected by the Matra line. Characters in a word may be connected by a Matra line. Here Matra line is detected first then vertical scanning is applied from the row which is just below the Matra line to the lower boundary line. Both procedures are shown in the Fig. 3.10 respectively.

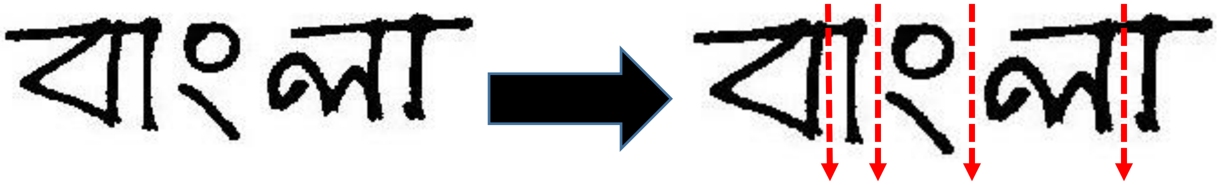


Fig. 3.10: Character Segmentation

3.5.5 Scaling

Bangla script has various sized characters. The characters need to be uniformed before extracting features for more appropriate results. Each character can be scaled into different sizes and shapes. To extract feature from it, it must be scaled into a standard size. Depending on the height and width of stored image, size of scanned image is scaled.

3.6 Feature Extraction

Feature extraction is done after the preprocessing phase in character recognition system. Feature extraction stage is to remove redundancy from data. Before building the feature extraction procedure, there are two important problems must be clarified which are feature extraction and feature selection. Feature extraction is related with which technique will be used to extract features from the image character as representations. On the other hand, in feature selection, the most relevant features to improve the classification accuracy must be searched. Generally there are two kinds of features.

They are as follows:

- Statistical features
- Structural features

3.6.1 Statistical Features

Representation of a document image by statistical distribution of points takes care of style variations to some extent. Although this type of representation does not allow the reconstruction of the original image, it is used for reducing the dimension of the feature set providing high speed and low complexity. The major statistical features mentioned below are used for character representation.

- **Zoning**

The frame containing the character is divided into several overlapping or non-overlapping zones. The densities of the points or some features in different regions are analyzed.

- **Crossings and Distances**

A popular statistical feature is the number of crossing of a contour by a line segment in a specified direction. The character frame is partitioned into a set of regions in various directions and then features of each region are extracted.

- **Projections**

Characters can be represented by projecting the pixel gray values onto lines in various directions. This representation creates one-dimensional signal from a two dimensional image, which can be used to represent the character image. There are some other statistical feature extraction methods also.

3.6.2 Structural Features

Structural features are based on topological and geometrical properties of the character. Various global and local properties of characters can be represented by geometrical and topological features with high tolerance to distortions and style variations. This type of representation may also, encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object. Various topological and geometrical representations can be grouped in four categories:

- **Extracting and Counting Topological Structures**

In this category, lines, curves, splines, extreme points, maxima and minima, cups above and below a threshold, openings, to the right, left, up and down, cross (X) points, branch (T) points, line ends (J), loops (O), direction of a stroke from a special point, inflection between two points, isolated dots, a bend between two points, horizontal curves at top or bottom, straight strokes between two points, ascending, descending and middle strokes and relations among the stroke that make up a character are considered as. Avoid combining SI and CGS units. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

- **Measuring and Approximating the Geometrical Properties**

In this category, the characters are represented by the measurement of the geometrical quantities such as, the ratio between width and height of the bounding box of a character, the relative distance between the last point and the last y-min, the relative horizontal and vertical distances between first and last points, distance between two points, comparative lengths between two strokes, width of a stroke, upper and lower masses of words, word length curvature or change in the curvature.

- **Coding**

One of the most popular coding schemes is Freeman's chain code. This coding is essentially obtained by mapping the strokes of a character into a 2-dimensional parameter space, which is made up of codes. There are many versions of chain coding. The character frame is divided to left-right sliding window and each region is coded by the chain code.

- **Graphs and Trees**

Words or characters are first partitioned into a set of topological primitives, such as strokes, holes, cross points etc. Then, these primitives are represented using attributed or relational graphs. Image is represented either by graphs coordinates of the character shape or by an abstract representation with nodes corresponding to the strokes and edges corresponding to the relationships between the strokes. Trees can also be used to represent the words or characters with a set of features, which has a hierarchical relation.

3.7 Recognition

OCR systems follow procedures for assigning an unspecified symbol to a predefined class that is known as recognition. There are various techniques to recognize a symbol. Some of them are as follows:

- **Template Matching**

Template Matching is the easiest method of character recognition. It performs matching between the predefine symbols and the character or word to be recognized. It compares the feature vector of predefined dataset with the feature vector of input symbol. And this matching procedure discovers the level of resemblance between these two vectors (structure, set of pixels, curvature etc.) for applying this technique, firstly we need two components first is binary input character and second is predefined dataset. This technique is flexible and straight forward. Similarities and dissimilarities between a gray-level or binary input symbol and a standard set of predefined prototypes are estimated by this approach. This estimation is done by matching operation corresponding to a similarity measure and match strength, a template matcher assign the input symbol to a class. The calculated recognition rate of this approach is very sensitive to distortion and image restoration. Some modification was performed in template matching for getting improvement in the result of classification. After modification, some other approaches were found, called Deformable Templates and Elastic Matching.

- **Statistical Techniques**

Statistical decision approach is related to statistical decision consequences and a group of optimality measures, which increases the occurrence of the perceived shape given the structure of a certain category. These procedures are depend on some hypothesis. These are:

- Distribution of feature group is uniform in the trounce instance, otherwise Gaussian,
- There are adequate statistics accessible for every category,
- Given set of images is allowed to take out a group of attributes which shows every different category of symbols.

A symbol contains various features. And feature extraction step of OCR provides feature vector for input image. And in classification step each feature of input image is compare with the feature vector of training dataset. This measurements from n-features of input image can be used to show n-dimensional feature vector volume and a vector whose correlation represents original character or word unit. There are some significant statistical methods, which are beneficial for getting good result of OCR. They are Likelihood or Bayes classifier, Hidden Markov Modelling (HMM), Quadratic classifier, Nearest Neighbour (NN), Clustering Analysis, Fuzzy Set Reasoning.

- **Neural Networks**

As human read the paragraph on his computer area, his eyes and brain bring OCR without his even attention. His eyes are identifying the shape of light and dark, those build the symbols (numbers, characters and word) printed on the area and his brain is using those to understand what I want to indicate (sometimes via taking single symbols or characters yet mainly via figure out whole letters and complete sets of letters suddenly).Problem with Character identification is belonging to heuristic argumentation as person can identify symbol, characters, and records by their learning, knowledge and practice. Therefore neural networks with more or less heuristic in essence are exceedingly acceptable for this type of issue. There are various categories of neural networks that are utilized for OCR recognition.

A neural network is a framework that contains of analogous interconnection of adjustable 'neural' processors. Neural network can execute calculations at an admirable rate juxtapose to the classical techniques due to its parallel nature. Neural network can accommodate for modification in the data and assimilate the attributes of input signal because of its adaptive nature. One node takes the output of previous one as input in the network and the final resolution based on the compound interaction of all vertices.

Artificial Neural Network (ANN) models involves of the following some components:

i. Topology –the process in which an ANN model is assembled into layers and arranges in a way in which these layers are interlinked;
ii. Learning –the procedure by which detail is accumulated in the grid; and
iii. Recall –how the accumulated detail is recovered from the grid.

The basic construction of an ANN model involve artificial neurons. The neurons are also called as nodes, processing elements (PEs), neurons, units, etc. And neurons are similar to biotic neurons in the person brain, which are assembled into layers (also called slabs). The simple ANN construction involves one or more than one hidden layers, an input layer, and an output layer.

- **Support Vector Machine Classifier**

The SVM technique was first arrived at erstwhile AT&T Bell Laboratories by Vapnik and his group. Support vector machine is also called two-class classifier. It works on distance between the classes [2]. This distance is called width of the margin. This distance is the separation to the closest training symbols. This is the free space throughout the decision boundary. Classification function is define by these training patterns. Theses training patterns called support vectors. Their number is decreased by increasing the perimeter. The support vectors exchange the originals with the main dissimilarity between support vector machine and an existing template matching procedures. They assign the categories by a decision boundary. This classifies use a linear differentiating hyperplane. This hyperplane increases the separation between two categories for creating a classifier. This classifier deals with linear and non-linear separation as well. The working of SVM is similar with the working of a statistical learning converter that merges points of distinct classes from n-dimensional area into a higher dimensional area where the two classes are more distinct. SVM works for finding a best hyperplane in high dimensional area so that it provides best separation for the two classes of points. The hyperplane is find by the points that are located nearest to the hyperplane, called support vectors. It is not necessary to have only one support vector, it may have more than one support vector on every side of the plane. The limitation of SVM classifier is that it classifies only two classes at a time. But practically, there are multiple classes to classify, so to classify multiple classes we require multiclass SVM. Multiclass SVM includes the construction of binary SVM classifiers for each pair of classes.

- **Combination Classifier**

A classification method has its own advantages and disadvantages. A classifier does not give best result for all classification problem. So sometimes, for solving classification problem, combination of classifiers required different classifiers follows different procedure so they are differ in their local and global performances. Each classifier, has in the feature space has its own area where it gives best result. Some classifiers like neural networks gives distinct outcomes for distinct initializations because of undirected inherent in the training process. Combination of various networks is use rather than selecting best network. So the advantages of all the attempts are taken out for solving classification problem. A set of classifiers contain different feature sets, different classification methods and different training sets. The output of all classifier is combined to improve overall classification accuracy. There are various combination schemes available. A difficult combination procedure involves a group of separate classifiers and a merger which combines the outcomes of the separate classifiers to give the final conclusion. Sometimes only two classifiers can solve classification problem, no need of third classifies. Sometimes three classifiers are required, sometimes more than three. So there are various methods to combine multiple classifiers. These methods are categories according to their structure. They are:

- Parallel,
- Hierarchical (tree-like) and
- Cascading (or serial combination)

Conclusion

This chapter represents a complete Optical Character Recognition system with a perspective of implementation. This segments emphasizes on the efficient ways involving line and word detection, zoning, character separations and character recognition. Employing Skewness correction, thinning and better approach in scaling have obviously enhanced the performance of the OCR system in comparison with the existing ones. Application of neural network in detection of characters has made the process even faster with optimum performance.