A Survey on Feature Extraction Techniques for Recognition of English Printed Characters

Harshad C. Bhuriya¹, Milin M. Patel²

¹ M.E Student, Dept. of C.E., SVIT, Vasad, Gujarat, India.
² Assistant Prof., Dept. of C.E., SVIT, Vasad, Gujarat, India.

Abstract—OCR stands for Optical Character Recognition, OCR is a generic process in area of intensive research field of artificial intelligence, pattern recognition and computer vision, aims to recognition of text from scanned printed document images, where data can be in machine printed format. OCR system consist of basically four stages which are pre-processing, segmentation, feature extraction and classification and followed by the actual recognition. The feature Extraction technique plays vital role in recognition of segmented printed character images. The work of offline character recognition of printed document images containing English Alphanumeric characters has been presented statistical, structural, directional feature extraction techniques have been applied over segmented character image.

Index Terms—OCR, SVM, PCA, ICR, BB.

I. INTRODUCTION

Optical Character Recognition is one of the most interesting and challenging research area in the field of image processing OCR aims at identifying segmented characters in images of printed text. Special applications and system include Printed document images, Bank check analysis and recognition and information retrieval, Forms Processing, OMR Sheet Processing.

The objective is to develop an offline OCR system to recognize Machine Printed Characters from the document images, which comprises of text in machine printed format, which would be convert editable form. An OCR system consist of different phases as Data Acquisition, Pre-Processing, Segmentation, Feature Extraction, Classification is shown in Figure 1. The feature Extraction technique plays vital role in recognition of segmented printed character images. The work of off-line character recognition of printed document images containing English Alphanumeric characters has been presented statistical, structural, directional feature extraction techniques have been applied over segmented character image.

II. PREVIOUS WORK

Rohit Verma, Dr. Jahid Ali [1] has identified by structural features which include concavities and convexities in the characters, Number of end points, No of holes in the character etc. Researchers have presented different feature extraction and classification techniques for recognition of characters. Nisha Sharma, Bhupendra Kumar, & Vandita Singh [3] have proposed Off-line character of recognition is applied to hand printed document images containing English Characters-Uppercase and Lowercase, Numerals and Special Characters. They applied to the different feature extraction techniques to segmented character images. Classification is by multilayer perception neural network with backpropogation & svm Classifier. The Recognition rate is achieved by using proposed system is 92.167%.

Sukalpa Chanda, Katrin Franke & Umapada Pal [5] in this paper researcher proposed a system for discriminating handwritten and printed text in the context of sparse data and arbitrary orientation. A chain-code feature is used with Support Vector Machine (SVM) classifier purposed for the handwritten and printed text recognition. This paper deals with discriminating hand-written and machine-printed text in Roman script with arbitrary orientation. Principal-component analysis (PCA) used to fix the orientation. And Characters in Roman printed text have a uniform shape. Whereas handwritten Roman text are of arbitrary curly styles. These characteristics are well represented by chain-code-histogram feature. The Recognition rate is achieved by using proposed system is 96.90%.
Gaurav Kumar, Pradeep Kumar Bhatia [6] has presented detailed review for different feature extraction techniques are used in image processing systems. Features like Statistical features, global transformation and series expansion, Geometrical and topological features.

Muhammad Arif Mohamad, Dewi Nsien, Hawadi Hassan, Habibollah Haron [7] have presented overview of handwritten character recognition and describes overview on feature extraction followed by current trend feature extraction and feature selection.

Lincoln Faria da Silva & Angel Sanchez [10] have presented Document image is firstly pre-processed by various techniques. Then the text is segmented at word level when each word is surrounded by a bounding box (BB). After word features are extracted from these BBs. The classification rules decide whether a BB contains printed or handwritten text. The main advantage of this, compared with other classifiers, is its accuracy, efficiency, simplicity and the low computation complexity. Classification is performed by words and not by line. The Recognition rate is achieved by using proposed system is 80%.

Abhishek Jindal & Mohd Amir [11] have presented Separate two types of texts and feed them to the respective engine - OCR (Optical Character Recognition) and ICR (Intelligent Character Recognition) engine to achieve optimal performance. This paper addresses the problem of classification of machine printed and handwritten text from acquired document images. Three features that are extracted are: Inter-character gap, Character Height and Baseline. An eight-neighbor connected component–labelling algorithm is used to get connected component information receives from ICR cell. The main advantage of this, compared with other classifiers, is its accuracy, efficiency, simplicity and the low computation complexity. The Recognition rate is achieved by using proposed system is 91%.

III. GENERAL OCR SYSTEM:

- **Image Acquisition:** The character recognition system acquires an input image through a digital scanner or any other suitable digital input device. The input captured may be in color, gray or binary from scanner or digital camera [14].

- **Pre-processing:** Pre-processing is the image Enhancement Technique. Pre-processing is important because it converts the image into a form suitable for subsequent processing and feature extraction. Major pre-processing steps are:
  1. Binarization
  2. Noise Removal
  3. Normalization
  4. Morphological Operations

  In character recognition systems most of the applications use gray or binary images since processing color images is computationally high. Image Binarization converts the image into bi-level form. The method used for Binarization is known as thresholding [14].

  Different thresholding techniques are global thresholding, local thresholding and Otsu’s thresholding. Noise is a random variation of image Intensity and visible as grains in the image. It may
produce at the time of capturing or image transmission. Common manifestation of noise in binary images takes the form of isolated pixels, salt-and-pepper noise. The major objective of noise removal is to remove any unwanted bit-patterns, which do not have any significance in the output. Major noise reduction technique is filtering. Mean filter, median filter and min-max filters are used for noise reduction. Median filter is widely used for the reduction of salt and pepper noise [14]. Normalization changes the range of pixel intensity values. Size Normalization is the process of converting the random sized image into standard sized image. This size normalization avoids inter class variation among characters [14]. Morphological operations which increase or decrease objects in size. Common morphological operations are erosion and dilation. Erosion shrinks the character image and dilation adds the pixels to the character boundary. Skeletonization or thinning is a morphological operation in which a single pixel wide representation of an image is obtained without changing its connectivity [14].

c) Segmentation: Segmentation is an important task of any character recognition system. It separates the image text documents into lines, words and characters. The accuracy of character recognition system mainly depends on the segmentation algorithm being used. First lines are segmented from the document and then each word is segmented from these lines and finally individual characters are extracted [14].

d) Feature Extraction: Different characters with different features can be identified by using OCR system; On the basis of these features of the characters are recognized. Features extraction can be considered as finding a set of features that define the shape of characters as precisely and uniquely as possible. There are number of feature extraction techniques used to recognize different English printed characters. Thus Feature Extraction can be defined as the process of extracting differentiating features from the matrices of digitized characters. A number of features have been found in literature on the basis of which the OCR system works to recognize the characters [1]. Feature Extraction methods are classified into four major groups as:

- Local or Structural Features
- Global or Statistical Features
- Directional Features
- Geometric and Regional Features

A. Local or Structural Features
Local method identifies local features of a character. Local features are based on topological and geometrical properties of the character [2]. Examples of local features are number of Horizontal lines or Vertical lines, number of End-points, number of cross points, Horizontal curves at top or bottom etc.

1) Number of End-points:
End-point defines starting and ending points of character. As shown in Figure 2, Character Y has 3 End-points [2].

![Figure 2: Character Y having 3 End-point](image)

2) End-Point Existence in Zone:
After finding numbers of endpoints, check in which zone endpoint exists. As shown in Figure 3, Y has 3 endpoints and endpoint existences are in Zone 1, Zone 3 and Zone 7 [2].

![Figure 3: End-point Existence in Zone 1, Zone3, and Zone 8](image)

3) Number of Vertical Lines:
Vertical Line feature describes that character contains vertical line. In English language vertical lines are always connected with some other part of the character and are not independent. As shown in Figure 4, character H contains 2 vertical lines [2].
4) Number of Horizontal Line:
Horizontal Line feature describes that character contains horizontal line. In English language horizontal lines are always connected with some other part of the character and are not independent. As shown in Figure 5, character E has 3 horizontal lines [2].

5) Geometrical features:
Geometrical features (Distance, angle) were obtained on each character image. Centre of the skeleton is calculated. Then image is divided into 3x3 zones as in Figure 6 [3]. In each zone Geometrical features are calculated as follows:

- Calculating average of distances of each pixel present in zone from center point as given by
  \[ D_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_m - n_k)^2 + (y_n - y_i)^2 \]  \( (1) \)
- Then average of angles of each pixel present in zone from centre point as given by 2 is calculated
  \[ A_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \tan^{-1} \left( \frac{y_n - y_i}{x_m - n_i} \right) \]  \( (2) \)
- Where \( x_m \) and \( y_n \) is the centre point co-ordinate of image, \( x_i \) and \( y_i \) are the coordinates of considered pixel and \( n_k \) is total number of data pixel present in the zone.

6) Topological Features:
Topological features (End points and Transitions) are extracted after each pre-processed image and deriving end points (E) with respect to the zone shown in Figure 3[2]. These features have been used to eliminate confusions among various characters at the stage of post-processing [3]. The number of transitions of black to white and white to black pixel horizontally and vertically were calculated for further rectifying the confusions at post-processing level. For example Vertical transition can be used to remove confusions among “y” and “z” shapes is shown in Figure 7 [3].

B. Directional Features:
A directional feature extraction technique-Chain Code Histogram (CCH) was implemented to obtain directional information in Fig. 7, based on contour of an image. This technique was also implemented at the zone level. There were 2 divisions (2x2 and 3x3), made for obtaining sub-images from the character image. For each sub-image, i.e. in each zone, the contour was traced to find the external boundary pixels and corresponding directions Fig. 8 were computed for each boundary line segment. Since the chain codes computed for each character was different, there would have aroused difficulty in comparison when used as features. Thus the chain code histogram for each segment was computed and used as feature. Hence, for each zone there was an addition of 8
features. For each character image having 13 zonal divisions, the total number of added features was 104(13x8) [3].

C. Global or Statistical Features
Global method identifies global features of a character. The global features are derived from the global distribution of pixels [2]. These features can be easily detected as compared to structural or local features. Statistical features are not affected too much by noise or distortions as compared to Structural features. A number of features are used for statistical features extraction; there are: Zoning, Projection Histograms, Crossings and distances, n-tupels. [2].

i. Zoning :
Zoning based feature extraction is one of the most popular methods. The character image is divided into predefined number of zones and a feature is computed from each of these zones. The character image is divided into several overlapping or non-overlapping zones. A character is usually divided into zones of predefined size. These predefined sizes are typically of the order 2×2, 3×3, 4×4 etc. In Figure 9, Character A is divided into 3×3 zones [2].

D. Geometric and Regional Features :
The paper [4] describes a geometry based technique for feature extraction applicable to segmentation-based character recognition systems. This method extracts the geometric features of the character contour. These features are based on the basic line types that form the character skeleton. The system gives a feature vector as its output. The method is implemented for English characters and the various steps involved in geometric method are:

(i) Initially pre-processing (Binarization, Skeletonization) is done on the input image.

(ii) Universe of discourse is defined as the shortest matrix that fits the entire character skeleton. The Universe of discourse is selected because the features extracted from the character image include the positions of different line segments in the character image.

(iii) After the universe of discourse is selected, the image is divided into windows of equal size, and the feature is done on Individual windows.

(iv) To extract different line segments in a particular zone, the entire skeleton in that zone should be traversed. For this purpose, certain pixels in the character skeleton were defined as starters, intersections and minor starters.

(v) After the line type of each segment is determined, feature vector is formed based on this information. Every zone has a feature vector corresponding to it. Under the algorithm proposed, every zone has a feature vector with a length of 8. The contents of each zone feature vector are

1) Number of horizontal lines
2) Number of vertical lines
3) Number of Right diagonal lines
4) Number of Left diagonal lines
5) Normalized Length of all horizontal lines
6) Normalized Length of all vertical lines
7) Normalized Length of all right diagonal lines
8) Normalized Length of all left diagonal lines
9) Normalized Area of the Skeleton

The number of any particular line type is normalized using the following method,
Value = 1 - ((number of lines/10) x 2)

Normalized length of any particular line type is found using the following method,

Length = (Total Pixels in that line type) / (Total zone pixels)

The feature vector explained here is extracted individually for each zone. So if there are N zones, there will be 9N elements in feature vector for each zone. For the system proposed, the original image was first zoned into 9 zones by dividing the image matrix. The features were then extracted for each zone. Again the original image was divided into 3 zones by dividing in the horizontal direction. Then features were extracted for each such zone. After zonal feature extraction, certain features were extracted for the entire image based on the regional properties namely

- Euler Number: It is defined as the difference of Number of Objects and Number of holes in the image. For instance, a perfectly drawn ‘A’ would have euler number as zero, since number of objects is 1 and number of holes is 2, whereas ‘B’ would have euler number as -1, since it have two holes.
- Regional Area: It is defined as the ratio of the number of the pixels in the skeleton to the total number of pixels in the image.
- Eccentricity: It is defined as the eccentricity of the smallest ellipse that fits the skeleton of the image [4].

e) Classification:

1) K-Nearest Neighbour Algorithm:

The k-Nearest Neighbors algorithm (k-NN) is a non-parametric method used for classification. The input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to class of that single nearest neighbor [13]. The idea behind k-Nearest Neighbor algorithm is quite straightforward. To classify a new character, the system finds the k nearest neighbors among the training datasets, and uses the categories of the k nearest neighbors to weight the category candidates [13].

2) Neural Network Algorithm

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems [20]. A neural network is a powerful data modeling tool that is able to capture and represent complex input/output relationships. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain.

Neural Network consists three layers - input layer, hidden layer (optional) and output layer.

1. Input Layer

The input layer is the conduit through which the external environment presents a pattern to the neural network. Input layer take input from the external world and encode it into a convenient form. Every input neuron should represent some independent variable that has an influence over the output of the neural network.

The number of input neurons for the OCR is the number of pixels that might represent any given character. A character which represents by a 5*7 grids has 35 pixels. So it has 35 input neurons.

2. Hidden Layer

Hidden layer can’t see or act upon the outside world directly. These inter-neurons communicate only with other neurons. Deciding the number of neurons in the hidden layers is a very important part of deciding your overall neural network architecture. Both the number of hidden layers and the number of neurons in each of these hidden layers must be carefully considered [13].

3. Output Layer

The output layer of the neural network is what actually presents a pattern to the external environment. The number of output neurons should be directly related to the type of work that the neural network is to perform. The number of output neurons used by the OCR program will vary depending on how many characters the program has been trained to recognize. The default training file that is provided with the OCR program is
used to train it to recognize 26 characters. Using this file, the neural network will have 26 output neurons. Neural network algorithm identifies the character by training the neural network. Feed forward Neural Network, Feedback neural network and Self Organizing Map are the types of neural network [13].

3) Support Vector Machine Algorithm

Support vector machines (SVMs also support vector networks) are a set of related supervised learning methods used for classification. SVMs are relatively new approach compared to other supervised classification algorithms, they are based on statistical learning theory developed by the Russian scientist Vladimir Naumovich Vapnik back in 1962 and since then, his original ideas have been perfected by a series of new techniques and algorithms [13]. Support vector machines have proved to achieve good generalization performance with no prior knowledge of the data. The principle of an SVM is to map the input data onto a higher dimensional feature space nonlinearly related to the input space and determine a separating hyperplane with maximum margin between the two classes in the feature space. This approach, in general, guarantees that the larger the margin is the lower is the generalization error of the classifier [13]. SVM algorithm is robust, accurate and very effective even in cases where the number of training samples is small.

IV. CONCLUSION

In this paper we have presented a survey of feature extraction techniques for optical character recognition of general script of printed English characters. A lot’s of research has been done in this field. Still the work is going on to improve the accuracy of feature extraction.

REFERENCES


[4] Dinesh Dileep. ”A Feature Extraction Techniques Based on Character Geometry for Character Recognition”


consonants for offline character recognition”