Off-Line Multi-Font/Size Character Recognition: A Review

Miss. Jigisha N. Parkar1 Prof. Rahul G. Patel2

1PG Student 2Assistant Professor

1,2Department of Electronics & Communication Engineering

1,2Sankalchand Patel college of Engineering, Visnagar, Gujarat Technological University,
University of Gujarat (India)

Abstract— Character recognition has been one of the most interesting and challenging research areas in the recent years. Many researchers develop scripts based on the different approaches used for the design of character recognition system. This paper provides a survey and classification of various character recognition techniques and describes a technique for converting textual content from a paper document into machine readable form. More precisely Character recognition is process of detecting and recognizing characters from input image and converts it into ASCII or other equivalent machine editable form. There are various approaches used for the design of OCR system, but this paper provide a review of existing works in handwritten and typewritten recognition of characters from image irrespective of their position, size, and various font styles based on the different approaches such as a Template matching, Statistical approach, Structural approach and Neural Network.

Key words: Off-line Optical Character Recognition, Template Matching, Statistical approach, Structural approach, Neural Network

I. INTRODUCTION

Optical Character Recognition is the process of translating images of handwritten, typewritten, or printed text into a format understood by machines for the purpose of editing, indexing/searching, and a reduction in storage size[4]. Character recognition (CR) has been extensively studied in the last half century and progressed to a level sufficient to produce technology driven applications[6]. Optical character recognition (OCR) represents numerous techniques for conversion of visual data of text (images, videos, scans, etc.) into machine-encoded character information [1]. Performances of OCR systems vary depending on type of text of scanned documents. Generally, we can distinguish three types of text: typewritten, printed and handwritten text. Typewritten and printed text represents the combination of a finite number of different characters[1].Computer system equipped with such an OCR system can improve the speed of input operation and decrease some possible human errors. Recognition of printed characters is itself a challenging problem since there is a variation of the same character due to change of fonts or introduction of different types of noises. Difference in font and sizes makes recognition task difficult if pre-processing, feature extraction and recognition are not robust. Therefore, a good character recognition approach must eliminate the noise after reading binary image data, smooth the image for better recognition, extract features efficiently, train the system and classify patterns[2].

II. DATA ACQUISITION

The progress in automatic character recognition systems is evolved in two categories according to the mode of data acquisition:

1) On-line character recognition systems
2) Off-line character recognition systems

Off-line character recognition captures the data from paper through optical scanners or cameras whereas the on-line recognition systems utilize the digitizers which directly captures writing with the order of the strokes, speed, pen- up and pen- down information. The various phases of OCR technique are[5]:

1) Digitization
2) Pre-processing
3) Segmentation
4) Feature Extraction
5) Classification
6) Post processing

III. PRE-PROCESSING

A series of operations have to be performed during the pre-processing stages. The main objective of the pre-processing is to organise the information so that the subsequent CR task become simpler. It essentially enhances the image rendering it suitable for segmentation[3]. Pre-processing aims to produce data that are easy for the CR systems to operate accurately. The main objectives of pre-processing are

1) Noise reduction
2) Normalization of the data
3) Compression in the amount of information to be retained [6].

The normalization of the characters is essential because of different types of fonts, which results in several variations in the shapes and sizes. Therefore, bring about uniformity among the input characters, all of them should be made of the same size.

IV. SEGMENTATION

Segmentation is an important stage because the extent one can reach in separation of words, lines, or characters directly affects the recognition rate of the script. There are two types of segmentation: external segmentation, which is the isolation of various writing units, such as paragraphs, sentences, or words, and internal segmentation, which is the isolation of letters, especially in cursively written words[6]. Segmentation is the process in which each character is separated from the text and resized into a fixed sized template. A Particular size of the test character is necessary for the recognition process. So after segmenting each
character is resized to a particular size. Incorrectly segmented characters are not in the expected position or parts of them are missed, may affect the OCR recognition. The most common character segmentation algorithms are based on vertical projection, pitch estimation or character sizes, contour analysis. In the segmentation process each character is separated by searching background columns between the characters. Between two characters there are at least one or two columns containing white background. Before that the lines are separated by the fact that there is at least one single row which contains white background. The characters are segmented by the MATLAB functions regionprops, bwlevel, rectangle and imcrop. After the segmentation stage the image is ready for feature extraction[9].

V. FEATURE EXTRACTION
The feature extraction step selects and prepares data which is used by a classifier to achieve the recognition task. Feature extraction involves representing a handwriting text by a set of discriminative features. The feature representation is based on extraction of certain types of information from the image. A feature extraction method that proves to be successful in one application domain may turn out not to be very useful in another domain. The feature extraction step selects and prepares data which is used by a classifier to achieve the recognition task[3]. Selection of feature is probably the single most important factor in achieving high recognition performance. In most of the recognition systems, in order to avoid extra complexity and to increase the accuracy of the algorithms, a more compact and characteristic representation is required[6]. On the other hand, transformed features generated by feature extraction may provide a better discriminative ability than the best subset of given features, but these new features (a linear or a nonlinear combination of given features) may not have a clear physical meaning[7]. A number of techniques are used for feature extraction; some of these are: moments, zoning, projection histograms, n-tuples, crossings and distances.

VI. CLASSIFICATION
The classification stage is the decision making part of the recognition system. The performance of a classifier relies on the quality of the features. There are many existing Classical and soft computing techniques for handwriting identification. They are given as:

A. Classical Techniques:
1) Template matching
2) Statistical techniques
3) Structural techniques

B. Soft Computing Techniques:
1) Neural networks (nns)
2) Fuzzy- logic technique
3) Evolutionary computing techniques

VII. POST PROCESSING
Post-processing stage is the final stage of the recognition system. It prints the corresponding recognized characters in the structured text form.

VIII. OCR SYSTEM DESIGN: APPROACHES
The four best known approaches for pattern recognition are:
1) Template matching, 2) Statistical Classification, 3) Syntactic or Structural approach, and 4) Neural Networks[7].

A. Template Matching Approach:
One of the simplest and earliest approaches to pattern recognition is based on template matching. Matching is a generic operation in pattern recognition which is used to determine the similarity between two entities (points, curves, or shapes) of the same type[7]. A gray-level or binary input character is directly compared to a standard set of stored prototypes. According to a similarity measure (e.g., Euclidean, Mahalanobis, Jaccard, or Yule similarity measures, etc.), a prototype matching is done for recognition. The matching techniques can be as simple as one-to-one comparison or as complex as decision tree analysis in which only selected pixels are tested[6]. The matching techniques can be classified as: Deformable templates & elastic matching, relaxation matching, direct matching[5].

B. Statistical Approach:
The purpose of the statistical methods is to determine to which category the given pattern belongs. By making observations and measurement processes, a set of numbers is prepared, which is used to prepare a measurement vector. Statistical classifiers are automatically trainable. The k-NN rule is a non-parametric recognition method. This method compares an unknown pattern to a set of patterns that have been already labeled with class identities in the training stage. A pattern is identified to be of the class of pattern, to which it has the closest distance. Another common statistical method is to use Bayesian classification. A Bayesian classifier assigns a pattern to a class with the maximum a posteriori probability[5].

C. Syntactic or Structural Methods:
In many recognition problems involving complex patterns, it is more appropriate to adopt a hierarchical perspective where a pattern is viewed as being composed of simple sub patterns which are themselves built from yet simpler sub patterns . The simplest/elementary sub patterns to be recognized are called primitives and the given complex pattern is represented in terms of the interrelationships between these primitives. In syntactic pattern recognition, a formal analogy is drawn between the structure of patterns and the syntax of a language. The patterns are viewed as sentences belonging to a language, primitives are viewed as the alphabet of the language, and the sentences are generated according to a grammar. Thus, a large collection of complex patterns can be described by a small number of primitives and grammatical rules. The grammar for each pattern class must be inferred d from the available training samples[7].

D. Artificial Neural Networks:
A neural networks composed of inter connected elements called neurons. A neural network can trained itself automatically on the basis of examples and efficient tools for learning large databases. This approach is non-algorithmic and is trainable[5].The most commonly used
family of neural networks for pattern classification tasks is the feed-forward network, which includes multilayer perception and Radial-Basis Function (RBF) networks. These networks are organized into layers and have unidirectional connections between the layers. Another popular network is the Self-Organizing Map (SOM), or Kohonen-Networx, which is mainly used for data clustering and feature mapping. Currently, neural network mainly used include the multi-layer perception network, Linear neural network, Back Propagation neural network, Radial basis neural networks, Feedback neural networks, Competition neural networks and so on. Classifiers Based on Back Propagation neural network has become mature, which have a good approximation of nonlinear mapping ability. Artificial Neural Network has many advantages, such as self-organizing, self-studying, parallel and self-adaptive. It doesn’t rely on accurate mathematical model and can deal with complex problems[8]. Back Propagation neural network structure is a kind of three or more than three neuron network, including input layer, hidden layer and output layer. Full connection between the upper and lower, and no connections between neurons in one layer. In the forward propagation process, the input information from the input layer upon hidden layer processing, and transmitted to the output layer; if the output level cannot get the desired output value, then turn back-propagation, the error signal along the original connection path back through the layers of neurons to modify the weights, so that the error signal is smallest[8].

IX. CONCLUSION
The success of OCR system quite heavily depended on the accuracy of the features extraction and classification techniques, so by properly choosing such type of techniques we can also obtain high accuracy. Template matching technique makes the procedure very simple, and the complexity of character shape is irrelevant, but, it suffers from the sensitivity to noise and is not adaptive to variations in writing style. An Neural Network is defined as a computing architecture that consists of a massively parallel interconnection of adaptive “neural” processors. Because of its parallel nature, it can perform computations at a higher rate compared to the classical techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. The advantages of using a neural network are its low noise-sensitivity, low classification time and high classification accuracy. A lot of research has been done in this field. Still the work is going on to improve the accuracy of feature extraction and classification techniques.

X. ACKNOWLEDGMENT
I would like to thank Prof. Rahul G. Patel sir for his guidelines in making this paper.

REFERENCES


