A Survey Paper on Character Recognition

Rinku Patel[1] Avani Dave[2]
[1]M. E. Student [2]Assistant Professor
Computer Engineering Department
[1, 2]L. J. Engg College, Ahmedabad, Gujarat, India

Abstract— Nowadays character recognition has gained lot of attention in the field of pattern recognition due to its application in various fields. It is one of the most successful applications of automatic pattern recognition. Research in OCR is popular for its application potential in banks, post offices, office automation etc. HCR is useful in cheque processing in banks; almost all kind of form processing systems, handwritten postal address resolution and many more. This paper presents a simple and efficient approach for the implementation of OCR and translation of scanned images of printed text into machine-encoded text. It makes use of different image analysis phases followed by image detection via pre-processing and post-processing. This paper also describes scanning the entire document (same as the segmentation in our case) and recognizing individual characters from image irrespective of their position, size and various font styles and it deals with recognition of the symbols from English language, which is internationally accepted.

Key words: Optical Character Recognition, HCR, Pattern Recognition, Off-Line Character Recognition, Template Matching.

I. INTRODUCTION

Optical Character Recognition is one of the most important gifts given by computer science to our mankind. It has made a lot of tedious work easy and speedy [11]. It is one of the most successful applications of automatic pattern recognition. Pattern recognition system classifies each member of the population on the basis of information contained in the feature vectors.

In Optical Character Recognition process [7], we convert printed document or scanned page to ASCII character that a computer can recognize. The image of document itself can be either machine printed or handwritten, or the combination of both. Computer system equipped with such an OCR system can improve the speed of input operation and reduce some possible human errors. Recognition of printed characters is itself a challenging task, 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. There may be noise pixels that are introduced due to scanning of the image. Besides, same font and size may also have bold face character as well as normal one. Thus, width of the stroke is also a factor that affects recognition [7]. Therefore, a good character recognition approach must eliminate the noise after reading binary image data, smooth the image for better recognition, extract features efficiently and classify patterns.

CR has been classified based upon the two important aspects: According to the manner in which data has been acquired (On-line and Off-line) and According to the text type (machine printed and handwritten). 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 [12].

II. DESIGN OF OCR

Various approaches used for the design of OCR systems are discussed below [7]:

A. Matrix Matching

Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

B. Fuzzy Logic

Fuzzy logic is a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black/white etc. An attempt is made to attribute a more humanlike way of logical thinking in the programming of computers. Fuzzy logic is used when answers do not have a distinct true or false value and there is uncertainly involved.

C. Feature Extraction

This method defines each character by the presence or absence of key features, International Journal of Computer Science & 92 Communication (IJCSC) including height, width, density, loops, lines, stems and other character traits. Feature extraction is a best approach for OCR of magazines, laser print and high quality images.

D. Structural Analysis

Structural Analysis identifies characters by examining their sub features- shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.

E. Neural Network

This strategy simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize characters through abstraction is great for faxed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns.

1) Structure of OCR Systems

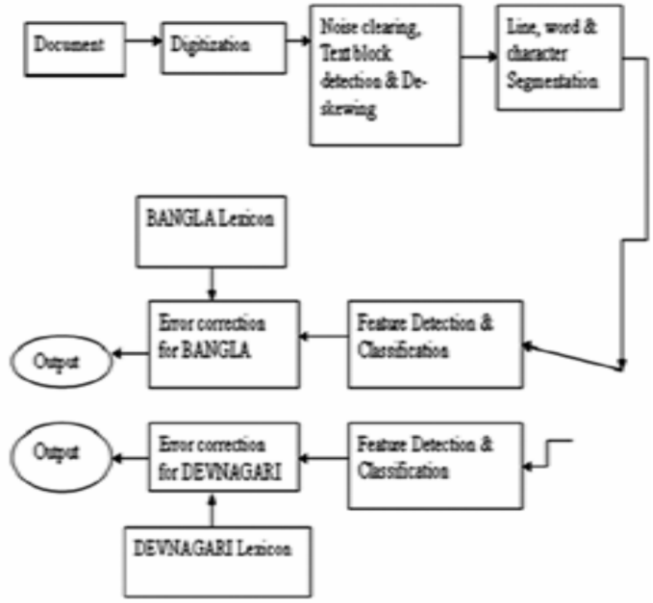

Fig. 1: Diagrammatic Structure of the OCR System (adapted from [13])

2) Stages in Design of OCR Systems

Various stages of OCR system design are given in figure 2.

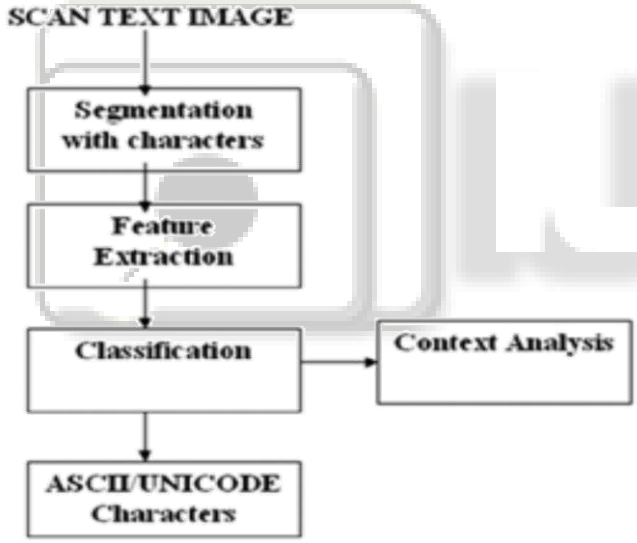

Fig. 2: Stages in OCR Design (adapted from [13])

III. PROPOSED OCR SYSTEM

The recognition of individual character from the image involves image to be analysed by different phases namely Image binarization, boundary detection, segmentation, thinning, image resize, template formation, template matching and classification. However apart from these phases other phases like noise removal can also be introduced. All the phases described below are useful to recognize only a single symbol from the image and scanning the entire document for recognition of set of symbols is described in Document Scanning section [2].

A. Binarization

The first step in Binarization is to convert a colour image Into a gray scale image. A gray scale digital image is an image in which each pixel is composed exclusively of shades of neutral gray, varying from black at the weakest intensity to white at the strongest intensity.

Algorithm: Binarization

1) Convert the colour image into gray scale image.
2) Find the gray level histogram of input image.
3) Find the valley point between two modes and select the point as threshold.

Then find the gray level histogram of gray scale image. The gray level histogram is composed of two dominant modes. One dominant mode corresponds to background and other is character image. Select the threshold value such that the two dominant modes meet at a valley point [9].

Colour image (AaZz) is converted into gray scale image and then into the binaries image as shown in Fig [3]:

Fig. 3: Colour, Gray scale and Binarized Image [1]

B. Boundary Detection

The boundary of character image can be produced by emphasizing regions containing abrupt dark-light transitions and de-emphasizing regions approximately homogenous intensity. In other words, by scanning the image pixel in both horizontal and vertical direction we can detect the boundary of the characters. In case of binary image character pixels are represented by black or '0' value and white pixels are by '1' thus by detection initiation of '0' value we can detect the boundary of the characters [1].

C. Segmentation

Segmentation is one of the most important phases of OCR system. By applying good segmentation techniques we can increase the performance of OCR. Segmentation subdivides an image into its constituent regions or objects. Basically in segmentation, we try to extract basic constituent of the script, which are certainly characters. This is needed because our classifier recognizes these characters only [7].

Overview of segmentation method based on vertical projection is presented as below:

In the method of vertical projection is dividing into two parts. First the whole image of the document is horizontally traced row by row to get the total number of lines present in the document. The presence or absence of black pixels is updated in row status vector. Then each line is taken under consideration and vertical tracing of each line is done separately to find total numbers of character presented in line. The present or absence of black pixels is updated in Column status vector [2].

Algorithm: Line detection from image

1) Start scanning the image horizontally from the topmost left corner row by row.
2) If any black pixel is encountered in a row make the row status as '0'.
3) If no black pixel in encountered in a row while tracing it then marks the row status as '1'.
4) By counting and following the total numbers of continuous '0' from row status vector number and position of lines can be obtained.
5) Algorithm: Character detection from the line
6) Take a single line under consideration.
7) Start scanning the image vertically from the topmost left corner column by column.
8) If any black pixel is encountered in a column mark the column status to '0'.
9) If no black pixel in encountered in a column while tracing it then marks the column status as '1'.
10) By counting and following the total numbers of continuous '0' from column status vector number and position of lines can be obtained.

D. Thinning

Thinning is a morphological operation which is used to remove selected foreground pixels from binary images. Thinning extracts the shape information of the characters. Thinning is also called skeletonization. It refers to the process of reducing the width of a line from many pixels to just single pixel. This process can remove irregularities in letters and in turn, makes the recognition algorithm simpler because they only have to operate on a character stroke, which is only one pixel wide as shown in Fig 4. It also reduces the memory space required for storing the information about the input characters and also reduces the processing time too [10].The algorithm structure is given below:

Algorithm: Thinning

1) Take the binary image of the isolated character under consideration.
2) While there are safe pixels for deletion, take out the pixel.
3) Repeat step 2 until all pixels are considered.

The safeness for deleting is checked on the basis of whether or not the pixel under consideration is border pixel. It also seen that, by deleting the pixel, the character loses one of the pixel of its skeleton, or not. If there is fear of losing skeleton information, the pixel is declared unsafe for deleting. This checking can be done by considering eight neighbours for each pixel (3x3 windows).

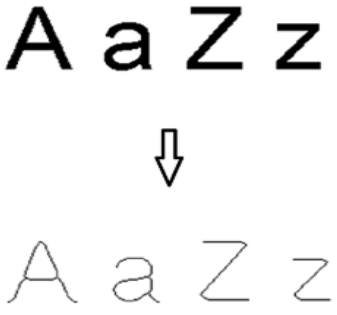

Fig. 4: Original and the thinned image

E. Resizing

Resizing of the thinned image is needed to be done to form character templates of fixed size. These character templates are used as reference database. Image resizing includes changing of the total size of image by increasing or decreasing number of pixels present in it. One of the commonly used image resizing technique is the interpolation method. Common interpolation algorithms can be grouped into two categories: adaptive and non-adaptive. Adaptive methods change depending on what they interpolate (sharp edges Vs. smooth texture), whereas non-adaptive methods treat all pixels equally. In our project we will only consider non-adaptive interpolation algorithms. The non-adaptive interpolation algorithms considered in our project are Nearest Neighbour Interpolation, Bilinear Interpolation, and Bicubic Interpolation [4].

F. Template Formation

The resized image of characters (alphabets A-Z & a-z) is further used for formation of character templates. These character templates are in form of feature vectors which are stored as reference data pattern. The reference data pattern is used at the time of template matching to the appropriate character [9].

IV. CLASSIFICATION

Classification basically decides the feature space to which the unknown pattern belongs. It is another most important component of OCR system. Classification is usually done by comparing the feature vectors corresponding to the input character with the representative of each character class. But before doing this the classifier should possess a number of training patterns. A number of classification methods were purposed by different researchers some of these are statistical methods, syntactic methods, template matching, artificial neural networks, kernel methods.

1) Statistical methods [14] Thema in purpose of the statistical methods is to determine to which category the given pattern belongs. To prepare a measurement vector a set of numbers is prepared, by making observations and measurement processes. Statistical classifiers are automatically trainable. The k-NN rule is a non-parametric recognition method. This method compares an unknown pattern with a set of patterns that have been already labelled 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. Bayesian classification is another common statistical method to be used. A Bayesian classifier assigns a pattern to a class with the maximum a posteriori probability. Other methods besides above methods are: Euclidean distance, Regularized Discriminant Analysis (RDA), Quadratic Discriminant Function (QDF), Mahanalobis distance, Linear Discriminant Function (LDF), cross correlation.

2) Syntactic or structural methods [14] Syntactic methods are better for classifying hand written texts. This type of classifier, classifies the input patterns on the basis of components of the characters and the relationship among these components. The primitives of the character are identified first and then strings of the primitives are checked on the basis of rules which are pre-decided. A character is

generally represented as a production rules structure, whose right-hand side represents string of primitives and whose left-hand side represents character labels. The right-hand side of rules is compared with the string of primitives extracted from a word. So classifying a character means finding a path to a leaf.

A. Template Matching

In our case we have done template matching by tracing the character template from all four directions i.e. from top, bottom, left and right. The distance between the border and character boundary of each and every row and column of both test and reference templates is measured and Euclidean difference between the distances is calculated. Such matching is done for each test pattern with every reference pattern. Grand total of the Euclidean difference from all four directions is calculated. The feature vector associated with the minimum grand total is declared as 'the best matched template'. This template is further used for classification and recognition of character [3] [5] [6].

In template matching, individual image pixels are used as features. Classification is performed by comparing an input character image with a set of templates from each character class. Each comparison results in a similarity measure between the input character and the template. One measure increases the amount of similarity when a pixel in the observed character is identical to the same pixel in the template image. If the pixels differ the measure of similarity may be decreased. After all templates have been compared with the observed character image, the character's identity is assigned as the identity of the most similar template [8]. In the recognition stage, a similarity or dissimilarity measure between each template Tj and the character image Z is computed. A common dissimilarity measure in the mean square distance (dj) and city block (d4) distance.

$$Dj = (Z(xi, yi) - Tj(xi, yi))\ 2$$

$$D4 = (|Z(xi) - T(xi)| + |Z(xj) - T(xj)|)$$

Where it assumed that the template and the input character image of the same size and sum is taken over the m pixels in the image.

Algorithm: Template Matching

1) Present the input image of any size and normalize to 50x30 pixels size. Find the Gray level histogram of input image.
2) Calculate the mean square distance, City block distance and hamming distances for the input image and strobe patterns.
3) Compare the dissimilarity between input image and template pattern.
4) If the dissimilarity below specified threshold, than it is called as Best Matched Template.

1) Artificial neural networks [14]

A neural networks composed of inter connected elements called neurons. A neural network can get itself trained automatically on the basis of efficient tools for learning large databases and examples. This approach is non-algorithmic and is trainable. Feed-forward networks the most commonly used family of neural networks for pattern classification task, which includes Radial-Basis Function (RBF) and multilayer perception networks. The other neural networks which are used for classification purpose are Vector Quantization (VQ) networks, Learning Vector Quantization (LVQ), Convolutional Neural Network, auto-association networks. But the limitation of all the systems based on neural networks is their poor capability for generality.

2) Kernel methods [14] some of the most important Kernel methods are Kernel Fisher Discriminant Analysis (KFDA), Kernel Principal Component Analysis (KPCA), Support Vector Machines, etc. Support vector machines (SVM) are a group of supervised learning methods which can be applied for classification. During a classification task usually data is divided into testing and training sets. The purpose of SVM is to produce a model, which predicts the target values of the test data. Different types of kernel functions of SVM are: Gaussian Radial Basis Function (RBF), Sigmoid, Linear kernel and Polynomial kernel.

V. CONCLUSION

In this paper, we have presented a survey of feature extraction and classification techniques for optical character recognition of general scripts. A lot of research has already been done in this field. Still the work is going on to further improve the accuracy of feature extraction and classification techniques. However the different methods of feature extraction and classification discussed here are very effective and useful for new researchers. Literature review shows that neural network is the prime choice among researchers for training purpose. However various kind of changes have been proposed in feature extraction methods. Work can be extended from single character or set of characters to document processing. Many applications are awaiting the enhancement in character recognition to be adopted it fully. Hybrid model can be proposed which counts on more than one feature extraction methods to discriminate characters properly.

REFERENCES

[1] Optical character recognition http://en.wikipedia.org/wiki/Optical_character_recognition
[2] Ritesh Kapoor, Sonia Gupta, C.M. Sharma, "Multi-Font/Size Character Recognition and Document Scanning" International Journal of Computer Applications, volume 23-No.1, June 2011
[3] Nafiz Arica, Fatos T., and Yarman-Vural, "An Overview Of Character Recognition Focused On Off-line Handwriting"
[4] http://www.cambridgeincolour.com/tutorials/image-interpolation.htm
[5] Mohammad Abu Obaida, Md. Jakir Hossain, Momotaz Begum, Md. Shahin Alam," Multilingual OCR (MOCR): An Approach to Classify Words to Languages" International Journal of Computer Applications (0975 – 8887) Volume 32– No.1, October 2011
[6] J. Pradeep, E. Srinivasan, S. Himavathi, "Diagonal Based Feature Extraction for Handwritten Alphabets Recognition System Using Neural Network "International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[7] Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash Yadav, "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network" International Journal of Computer Science and Communication, Vol. 1, No. 1 January-June 2010,pp.91-95
[8] Jayashree R. Prasad, Dr. U. V. Kulkarni, Rajesh S. Prasad, "Template matching algorithm for Gujarati Character Recognition" Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09
[9] D. A. Jadhav and G. K. Veeresh, "Multi-Font/Size Character Recognition" International Journal of Advances in Engineering & Technology, May 2012, Vol. 3, Issue 2, pp. 442-445
[10] B. Indira, M. Shalini, M. V. Ramana Murthy, Mahaboob Sharief Shaik, "Classification and Recognition of Printed Hindi Characters Using Artificial Neural Networks" I.J. Image, Graphics and Signal Processing, July 2012,6 15-21
[11] Rohit Verma and Dr. Jahid Ali, "A-Survey of Feature Extraction and Classification Techniques in OCR Systems" International Journal of Computer Applications & Information Technology, Vol. I, Issue III, November 2012 (ISSN: 2278-7720)
[12] Shabana Mehfuz, Gauri Katiyar, "Intelligent Systems for Off-Line Handwritten Character Recognition: A Review" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 4, April 2012).
[13] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review."(2012): 1-1.
[14] Rohit Verma, Dr. Jahid Ali" A-Survey of Feature Extraction and Classification Techniques in OCR Systems" International Journal of Computer Applications & Information Technology Vol. I, Issue III, November 2012 (ISSN: 2278-7720)

