International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 2 | ISSN : 2456-3307

Optical Character Recognition in Devnagri Script

Naeem Sunesara, Tanmay Bane, Dipesh Pawar, Nikhil Saggam

Department of Information Technology, Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Mumbai, Maharashtra, India

ABSTRACT

This is the software which will recognize the characters from online or offline document (in image format) and use it as individual user profile. Here, the software OCR will recognize Devnagri characters. OCR is an Optical character recognition and is the mechanical or electronic translation of images of handwritten or typewritten text (usually captured by a scanner) into machine-editable text. OCR is a field of research in pattern recognition, Artificial Neural Networks and Kohonen Network.

Keywords: OCR, Preprocessing, Segmentation, Feature Extraction, Classification, Kohonen Algorithm

I. INTRODUCTION

Optical Character Recognition is a process by which we convert printed document or scanned page to ASCII character that a computer can recognize. 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 preprocessing, 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. 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. Till now there is no complete OCR for printed Devnagari Script which gives 100% success rate.

II. METHODS AND MATERIAL

1. Existing System

Older systems involve using entire image for recognition thus adding complexities whereas this system involves dividing the image into small blocks. Existing systems are based on assumptions which may or may be correct but this method uses chain code algorithm which gives the accurate results. Every feature vector in the earlier systems needs to be labelled and trained thus increasing time and space complexities but this system uses chain code algorithm so this problem is solved. Existing OCR systems generally show poor performance for documents like old books: print and paper quality inferior due to aging, Copied Materials: documents like photocopies or faxed documents, where print quality is less blurring than linear smoothing filters of similar size.

2. Proposed System

Our proposed system is a character recognition system that supports recognition of the characters of Devnagri letters and numerals. This feature is what we call feature extraction which eliminates the problem of heterogeneous character recognition and supports multiple functionalities to be performed on the image.
Following steps have been followed in the design of proposed OCR system:

- Preprocessing
- Segmentation
- Feature Extraction
- Classification

CSEIT1722356 | Received : 18 April 2017 | Accepted : 28 April 2017 | March-April-2017 [(2)2: 1075-1077]

2.1. Preprocessing

In the proposed OCR system, text digitization is done by a flatbed scanner having resolution between 100 and 600 dpi. The digitized images are usually in gray tone, and for a clear document, a simple histogram based threshold approach is sufficient for converting them to two tone images. The histogram of gray values of the pixels shows two prominent peaks, and a middle gray value located between the peaks is a good choice for threshold. For salt and pepper noise we generally use median filter. Median filter replaces the value of a pixel by the median of gray levels in the neighborhood of that pixel (the original value of the pixel is included in the computation of the median), Median filters provide excellent noise reduction capabilities, with considering less blurring than linear smoothing filters of similar size as shown in figure 1 & 2.

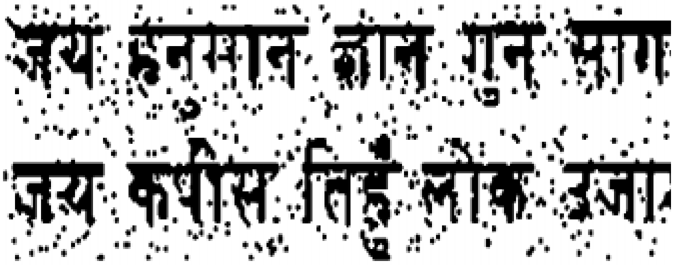

Fig 1: Image with Salt and Pepper Noise

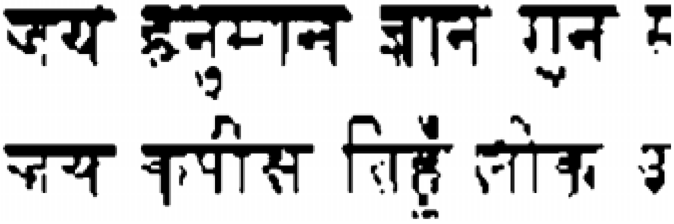

Fig 2: Image without Salt and Pepper Noise

Derivative operator enhances edges and other discontinuities (noise) and deemphasizes area with slowly varying gray level values.

2.2. 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. Segmentation phase is also crucial in contributing to this error due to touching characters, which the classifier cannot properly tackle. Even in good quality documents, some adjacent characters touch each other due to inappropriate scanning resolution.

2.3.Feature Extraction

In this phase, features of individual character are extracted. The performance of an each character recognition system that depends on the features that are extracted. The extracted features from input character should allow classification of a character in a unique way. We used diagonal features, intersection and open end points features, transition features, zoning features, directional features, parabola curve fitting–based features, and power curve fitting–based features in order to find the feature set for a given character.

2.4. Classification

Classification is performed based on the extracted features. Here we are using Artificial Neural Network approach. For initial classification of characters, we consider three features as follows:

- Mean Distance;
- Histogram of projection based on spatial position
- of pixel;
- Histogram of projection based on pixel value.

Artificial Neural Network approach has been used for classification and recognition. It is a computational model widely used in situation where the problem is complex and data is subject to statistical variation. Training and recognition phase of the ANN has been performed using conventional back propagation algorithm with two hidden layers. The architecture of a neural network determines how a neural network transfers its input into output. This transfer can be viewed as a computation

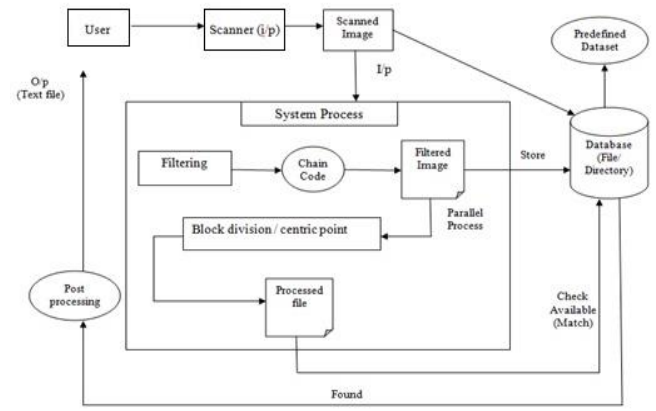

Figure 3: Architecture Diagram

Advantages of Proposed System

1. Data entry for business documents
2. Automatic insurance documents key information extractions
3. Extracting business card information into contact list
4. More quickly make textual versions of printed documents

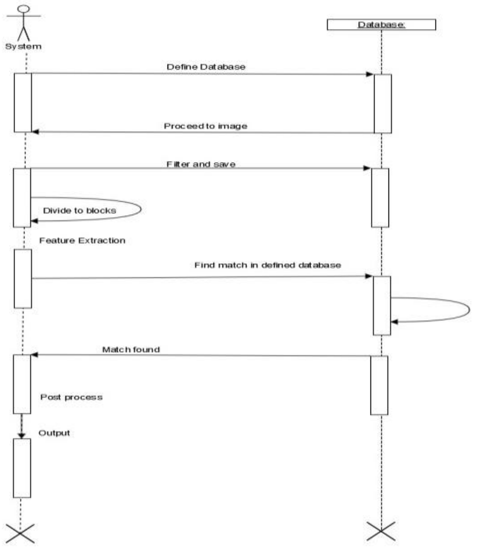

Figure 4: Sequence Diagram of the Proposed System

III. RESULTS AND DISCUSSION

Implementation and Design

The method adopted for recognition is a rather simplistic approach to image analysis. Since the number of probable candidate has already been reduced to a small value, the recognition algorithm need not be a very complex one. We compare the attributes of the portion to the left of the vertical bar to a two standard templates, one representing the left portion of ka and the other representing the left portion of fa. The one that gives a better match is selected. To take into account the variations possible, five standard templates of each character are taken, and their mean is calculated to give the ideal template.

IV. CONCLUSION AND FUTURE SCOPE

Given enough entrepreneurial designers and sufficient research and development dollars, handwritten text image recognition can become a powerful tool for future data entry applications. However, the limited availability of funds in a capital-short environment could restrict the growth of this technology. But, given the proper impetus and encouragement, a lot of benefits can be provided by this system. The automated entry of data is one of the most attractive, labor reducing technology. The recognition of new font characters by the system is very easy and quick. We can edit the information of the documents more conveniently and we can reuse the edited information as and when required. The extension to software other than editing and searching is topic for future works.

V. REFERENCES

[1]. S. Morietal, "Historical Review of OCR Research and Development", Proceeding IEEE, 80, no 7, pp. 1029-1058, July 1992

[2]. Sushree Sangita Patnaik and Anup Kumar Panda Particle Swarm Optimization and Bacterial Foraging Optimization Techniques for Optimal Current Harmonic Mitigation by Employing Active Power Filter Applied Computational Intelligence and Soft Computing Volume 2012, Article ID 897127.

[3]. Dileep Kumar Patel, Tanmoy Som1, Sushil Kumar Yadav, Manoj Kumar Singh," Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric" JSIP 2012, 208-214

[4]. P. Wei Sch. of Electr. Inf., Zhongyuan Univ. of Technol., Zhengzhou, China Liang Zhang ; Changzheng Ma. "Fast median filtering algorithm based on FPGA median".

[5]. P. Dreuw, G. Heigold, and H. Ney. "Confidence- and margin-based mmi / mpe discriminative training for offline handwriting recognition." Anal. Recognition, 14(3):273–288, Sept. 2011