International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.6, No.9, September 2017 DOI:10.15693/ijaist/2017.v6i11.417-420

A Technique for Recognition of English Characters in Printed Document Images

Ravikumar B. Chawhan[1], A. Kiran Kumar Reddy[2]
[1,2] Assistant Professor, Department of CSE,
[1,2] Samskruthi College of Engineering & Technology, Ghatkesar,Telangana.
Email: [1]ravismsk1 @gmail.com, [2]kkiiran14@gmail.com

Abstract-- A new method for character recognition from printed English document images is presented. The method uses zone wise statistical features for recognition of characters. The method works in two phases namely: training and testing. During training, the training sample character images are processed to extract zone wise statistical features and stored into knowledgebase. The features stored in the knowledge base are fed to neural network for training. Further, during testing a sample character test image is processed for feature extraction and recognized using neural network classifier. The method is evaluated for 520 English character images from printed document images and overall recognition accuracy of 97.75% is achieved. The method is found to be efficient for variations in font style and size of the characters.

Keywords: character recognition, zone wise statistical features, training, testing.

I. INTRODUCTION

Optical Character Recognition is a process that recognizes printed or handwritten characters electronically. Recognition engine of the system will interpret the images and convert images of handwritten or printed characters to ASCII data i.e. machine readable characters. The work of computer in modern society is to process huge volumes of data. Large amount of handwritten and printed data has to be fed as an input to the computer for business requirements and economic reasons. The information that the human operator's type into the computer is present on the paper, examples includes checks, payment slips, income tax forms, billions of letters in the mail, and many business forms and documents. Hence, OCR lightens error prone and time consuming work by recognizing characters at high speed. In this work, a method for recognition of printed English characters in document images is presented. It works in two phases namely: training and testing. During training, the training sample character images are processed to extract zone wise statistical features and stored into knowledgebase. The features stored in the knowledge base are fed to neural network for training. Further, during testing a sample character test image is processed for feature extraction and recognized using neural network classifier. The method is evaluated for 520 English character images from printed document images and overall recognition accuracy of 97.75% is achieved. The method is found to be efficient for variations in font style and size of the characters.

II. RELATED WORK

A method for analysis of printed characters of high degradation is employed in [1]. They have proposed a coordinated framework system for handling of text characters in printed degraded documents. They consider ancient printed text for their study which used wavelet-based decomposition and filtering for noise removal. Features used are energy function that accounts for data consistency, smoothness and geometrical constraints and neural network classifier. Structural features for recognizing degraded printed gurmukhi script is presented in [2]. Based on printing quality of the document that is given as input the performance of OCR system is decided. Features employed are number & intersections between character & straight lines, holes and concave arcs, number & position of end points & junctions. KNN and SV classifier are used and recognition accuracy of 83.60% is achieved. A work on Compression and String Matching Method for Printed Document Images is discussed in [3]. Feature employed in this work is GDF (which calculates a gradient vector of pixel value at each position of a target image) generates feature based on distribution of gradient vector and string matching includes pseudo coding. A methodology for handwritten character recognition through two-stage foreground sub-sampling is reported in [4]. The work employs statistical and structural features and SVM classifier. A model for Statistical Machine Translation as a Language Model for Handwriting Recognition is given in [6]. A new type of language model used in addition to the standard n-gram LM. It automatically translate each OCR hypothesis into another language, then create a

feature score based on how "difficult" it was to perform the translation.

III. PROPOSED WORK

The method uses zonewise statistical features for the recognition of English characters in printed document images. The proposed method comprises of 4 phases: Preprocessing, Feature Extraction, Training and Testing. The block schematic diagram of the proposed model is given in Fig 1. The detailed description of each of processing step is presented in the following subsections:

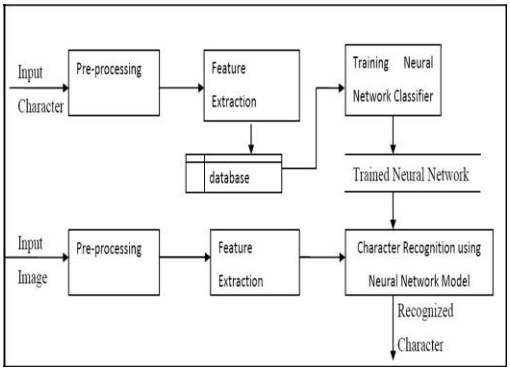

Fig 1. Block diagram of proposed model for character recognition

A. Preprocessing

Pre-processing enhances the visual appearance of the image. The purpose in this phase is to make input characters of standard size and ready for feature extraction. Many preprocessing methods are followed. This involves 3 stages of preprocessing. They are gray scale conversion, binarization and resizing of image.

B. Feature Extraction

In this step, the image is processed to extract the features using zoning process. The character image is divided into 36 zones (each zone is of size 5*5 as shown in figure 2.3) and features are computed from each of these zones. The sum of on pixels is determined as a feature value. From each character 36 features are been extracted and stored in the feature vector called *rsum1*. The features extracted from each character are stored in knowledgebase which is further used for training.

Algorithm

Input: Binary_Img (size 30*30)
Step 1: Calculate newMx, i.e image of 5*5 zones; which displays image in binary format.
Step 2: rsum is calculated as total number of white pixels in each row and is stored in rsum1 vector.
Step 3: Sum of each zone is computed and is stored in variable called as value.
Step 4: Value represents a single feature of every character.
Step 5: repeat step1 to step5 for all 36 zones and calculate 36 features of a single character
Step 6: store the 36 feature values in vector called rsumvector.
Step 7: repeat the above steps for every input image. Step 8: Store all the features of every character in the knowledgebase which is further used in recognition. Let I be image of size m*n blocks or zones, where size of each i^{th} block B_i is 5*5. The sum of on pixels in each B_i is defined as a feature value as given in equation (1).

$$f_i = \sum_{p=1}^{} \sum_{q=1} B_i(p,q) \quad (1)$$

The features are stored in the feature vector as given in (2).

$$FV = [\, f_1 , f_2 ..... f_n \,] \quad (2)$$

C. Training

The features obtained from the training samples are used to train the neural network model. The neural network used in the work consists of 36 input nodes, 20 hidden layers and 5 output nodes.

D. Recognition

In recognition, input image is presented to obtain features and are given as input to trained model, the trained model recognizes given sample and produces output and corresponding pattern as shown in figure 2.

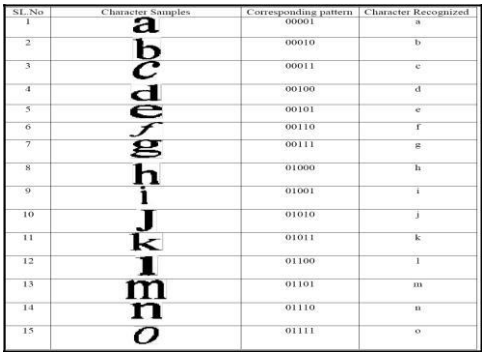

Fig 2. Recognition Samples

IV.EXPERIMENTAL RESULTS

The proposed methodology has been evaluated for 520 English character images found in printed documents. The method achieves recognition accuracy of 97.75%. The system is efficient and insensitive to the variations in size and font. The experimental results of testing various character images with varying font style and sizes is given below

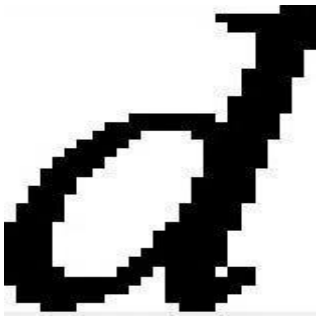

Fig 3. Input image

A sample character image from printed English document is given as input. The input image is uppercase character of English document. The character image is preprocessed for binarization of image. The binary image is shown in figure 4.

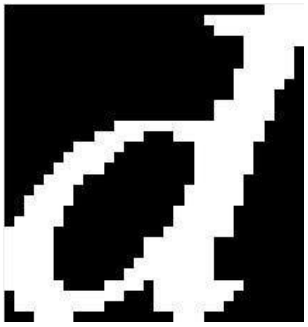

Fig 4. Binary image

The binary image is resized to standard image of size 30X30 pixels and zonewise profile features are computed and is stored in vector.

Table I. Feature Values

0	0	0	0	11	23
0	0	0	0	11	15
0	8	11	10	24	4
9	13	0	9	20	0
24	1	4	22	9	0
20	13	9	23	10	0

The features stored in knowledgebase are used for training the neural network. The test image is processed for feature extraction and recognized using neural network classifier and corresponding pattern is as follows

y2 = 0 0 1 0 0
dec = 4

V.SYSTEM PERFORMANCE ANALYSIS

The experiment is performed on 520 English character images. The following table shows overall performance of the system.

TABLE II. System Performance Analysis

Total number of images in dataset	UpperCase Characters		LowerCase Characters		Overall Recognition Accuracy
	Number of Characters recognized correctly	Accuracy of the method	Number of Characters recognized correctly	Accuracy of the method	
520	258	99%	251	96.5%	97.75%

CONCLUSION AND FUTURE WORK

A new method for character recognition from printed English document images is presented. The method uses zone wise statistical features for recognition of characters. The method works in two phases namely: training and testing. During training, the training sample character images are processed to extract zone wise statistical features and stored into knowledgebase. The features stored in the knowledge base are fed to neural network for training. Further, during testing a sample character test image is processed for feature extraction and recognized using neural network classifier. The method is evaluated for 520 English character images from printed document images and overall recognition accuracy of 97.75% is achieved. The method is found to be efficient for variations in font style and size of the characters.

The proposed method needs to be extended by employing new features to improve the performance of the system. The method can also be extended to provide the characters automatically from English printed documents to the recognition system.

REFERENCES

1) Anna Tonazzini, Stefano Vezzosi, Luigi Bedini, “Analysis and recognition of highly degraded printed characters”, IJDAR International Journal on Document Analysis and Recognition, Volume 6, Issue 4, pp. 236-247, 22 June 2003.
2) M. K. Jindal, R. K. Sharma and G. S. Lehal, “Structural Features for Recognizing Degraded Printed Gurmukhi Script”, in Proceedings of the IEEE 5th International Conference on Information Technology: New Generations (ITNG 2008), pp. 668-673, Published by IEEE Computer Society, April 2008(ISBN No. 978-0-7695-3099-4/08).
3) Hajime Imura and Yuzuru Tanaka, “Compression and String Matching Method for Printed Document Images”, 10th International Conference on Document Analysis and Recognition, pp. 291-295, 2009.

4) Georgios Vamvaka, Basilis Gatos, Stavros J. Perantonis, "Handwritten character recognition through two-stage foreground sub-sampling", pattern recognition 43, pp. 2807–2816, 23 February 2010.
5) Hai Guo, Jing-ying Zhao, "A Chinese Minority Script Recognition Method Based on Wavelet Feature and Modified KNN", Journal of software, vol. 5, no. 2, pp. 251-258, february 2010.
6) Jacob Devlin, Matin Kamali, Krishna Subramanian, Rohit Prasad and Prem Natarajan, "Statistical Machine Translation as a Language Model for Handwriting Recognition", International Conference on Frontiers in Handwriting Recognition, pp. 291-296, 2012.
7) Manoj Kumar Shukla, Haider Banka, "A Study of Different Kinds of Degradation in Printed Bangla Script",1st International Conf. on Recent Advances in Information Technology, ISM, Dhanbad, pp.119 – 123, march 2012.
8) Rakesh Rathi, Ravi Krishan Pandey & Vikas Chaturvedi, Mahesh Jangid, "Offline Handwritten Devanagari Vowels Recognition using KNN Classifier", International Journal of Computer Applications (0975 – 8887) ,Volume 49, No.23,pp. 11-16, July 2012.
9) Rajiv Kumar Nath, Mayuri Rastogi, "Improving Various Off-line Techniques used for Handwritten Character Recognition: a Review", International Journal of Computer Applications (0975 – 8887), Volume 49, No.18, July 2012.
10) Deepika Ghai, Neelu Jain, "Text Extraction from Document Images- A Review", International Journal of Computer Applications (0975 – 8887), volume 84, No 3, pp. 40-48, December 2013.
11) Sukhpreet Singh, "Optical Character Recognition Techniques: A Survey", Journal of Emerging Trends in Computing and Information Sciences, vol. 4, No. 6, pp. 545-550,ISSN 2079-8407, June 2013.
12) Gaurav Kumar, Pradeep Kumar Bhatia, "Neural Network based Approach for Recognition of Text Images", International Journal of Computer Applications (0975 – 8887) ,Volume 62, No.14, pp. 8-13, January 2013.
13) Arshiya Nain1, Sukhwinder Singh, "Character Recognition in Natural Images", International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463, vol. 3, Issue 3, pp. 123-126, March-2014.

Author Profile

Ravikumar B. Chawhan had his M. Tech. from Basaveshwar Eng. College Bagalkot, Karnataka. He is currently working in Samskruti College of Eng. & Technology Ghatkesar (Mandal), Telangana.

A. Kiran Kumar Reddy had his M.Tech from CMR Institute of Technology, Hyderabad. He is currently working in Samskruti College of Eng. & Technology Ghatkesar(Mandal), Telangana.