International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 02 | Feb -2017 www.irjet.net p-ISSN: 2395-0072

Design and Description of Feature Extraction Algorithm for Old English Font

Sreesha Bhaskar[1], Dr Saravanan K N [2]
[1] *Dept. of Computer Science, Christ University, Bangalore, India*
[2]*Associative Professor, Dept. Of Computer Science, Christ University, Bangalore, India*

Abstract - *The recognition of character has been challenging these days. In the field of text recognition of alphabets, much advancement has been done. This paper proposes a design and description of feature extraction algorithm for recognizing Old English Font. The proposed method consists of four stages, which states, data collection, preprocessing, feature extraction and recognition (minimum distance classifier). The old English font characters must be preprocessed in order to remove noises. The binarized image can be used for feature extraction method. Algorithm Old English Font is used for feature extraction. Minimum distance classifier has been used for the recognition of the character. The method gives a 79% satisfactory recognition rate.*

Key Words: **Old English font characters, Pre-processing, Feature extraction, Minimum distance classifier**.

1. INTRODUCTION

Character recognition or optical character recognition is a mechanism of conversion of handwritten or printed text (which is scanned) into machine editable form. The research on OCR system is improving day by day. The system uses OCR engine, a computer program which will recognize the character image represents. The OCR was proposed by Taushek and Handel. The proposed system focuses on old English font characters. Old English font is difficult for a human to read. There are few letters which are very difficult to recognize. The starting stage is scanning of the printed document in to image. The image which is scanned is processed in different stages and converts to character code, so that it can be edited and manipulated. Pre-processing is necessary after scanning a document since it contains many noises. Methods like binarization, noise removal, feature extraction, classification etc has been done. The type of pre-processing algorithms depends on the age of the document, quality of the paper and scanned image. Accuracy of OCR system depends of algorithms which we are using for feature extraction and classification. Though it does not mean other methods are not important. There is an increment found in use of multi-lingual documents which requires feature list to be more novel and classification process to be more precise for achieving more accuracy rate and less error rate in recognition process.

1.1 PROBLEM DEFINION

Old English or black letters were used thousand years ago and are important part of European cultural heritage.. Libraries, as the model preservers of printed archives, can no longer avoid the worldwide spin of digitization. Colleges, libraries wishing to digitize their old document collection face a great challenge. Recognition of old English font characters is an area where much researches has not done.

1.2 LITERATURE REVIEW OF EXISTING TECHNIQUES

Ankit et al. in [1] explains how to identify engine number which has been engraved in two wheeler and four wheeler using optical character recognition techniques. This paper gives an accurate rate of 99.9%. In this paper all the pre-processing steps are done using java programming. An accurate result of 99.9% is obtained by storing all the images in the database with a particular format. Using java programming, vehicle registration gives correct result which matches the images in the database. Engine number which written in any language and any font will be identified with highest accuracy. There is some limitation with this technique, as number of images in the database increases a good care has to be taken. Identifying engraved number during vehicle moving condition kept as a future work. Adityaraj in [2] explains about how feature extraction carried out using detection of vertical line a character and detection of open space in lower zone of a character. Classification is done by combination of binary tree and naive Bayesian classifier. Image enhancement has been done using spatial filtering. Binarization is done using OTSU's thresholding and for segmentation of each character is done using bounding box methodology. The feature extraction used till now are moment based feature and structural feature. Proposed OCR system uses two structural features. Detection of vertical line in a character is used to classify the oriya script into characters having vertical lines and characters with no vertical lines. Detection of open space in lower zone classifies characters which has open space on the lower zone. All the characters which are categorized in to four classes are put into the Bayesian classifier for the recognition process. It gives an average 99.25% of accuracy. Shuwair et al. in [3] explains the proposed OCR for Urdu framework which is built on MATLAB and Microsoft C#.Net.

The OCR was introduced for both offline and online (just for confined characters). The framework has tried on various printed and handwritten records with various text styles and scripts. In extracting the content lines it gives 97.09%. The framework was tried with 1050 characters which are single and ligatures in which 98.86% of accuracy was found during extraction. 97.12% of accuracy was found in recognition. The general effects of Urdu after OCR framework were very reassuring for offline and online. Future work incorporates with improvements of algorithm. Complete record investigations should be finished. Segmentation free approach required a lot of calculation so look into must be done on character segmentation. D. N. Hakro et al. in [4] explains the endeavours of analysts on Arabic and its related languages. The overview is sorted out in various areas, in which presentation is trailed by properties of Sindhi Language. OCR prepares strategies and techniques utilized by different specialists are presented**.** By developmental methods, languages like Sindhi and Arabic script's OCR issues can be settled and it is coordinated towards its development level like the Latin OCRs have accomplished. A high productivity OCR is the yield of enhancing ordinary procedures of division and highlight extraction. This paper also made some data regarding Arabic and its similar types of languages. Dr Mrs. V.V Patil et al. in [5] describes a varied framework for the traffic signs by a automatic recognition of characters. To describe the search region inside the image a particular scene structure is utilized, in which the traffic signs are then utilized. To locate the number of candidates, Maximally stable extremal regions (MSERs) and shade, immersion, and esteem shading thresholding are utilized, which are then diminished by applying confinements in light of transient and essential information. Single characters are identified as MSERs and are organized into different lines, before utilizing the optical character recognition (OCR) method. And the recognition accuracy is also improved. A novel framework for the programmed identification and acknowledgment of content in activity signs in light of HSV and MSERs thresholding has been introduced. Point of view amendment and worldly combination of competitor areas of content were utilized to enhance OCR comes about. Both the discovery and acknowledgment stages of the framework were approved through relative examination, accomplishing the Fmeasure of 0.93 for identification, 0.89 for acknowledgment, what's more, 0.87 for the whole framework.

2. DATA COLLECTION

In this work, the database consists of both capital and small letters for training and testing purpose. It has separate training and testing sets. 20,800 capital letters have been used only for training purpose and 9621 character samples are tested for testing purpose. In total 31,200 small letters are taken for training purpose and 12300 character samples are taken for testing purpose.

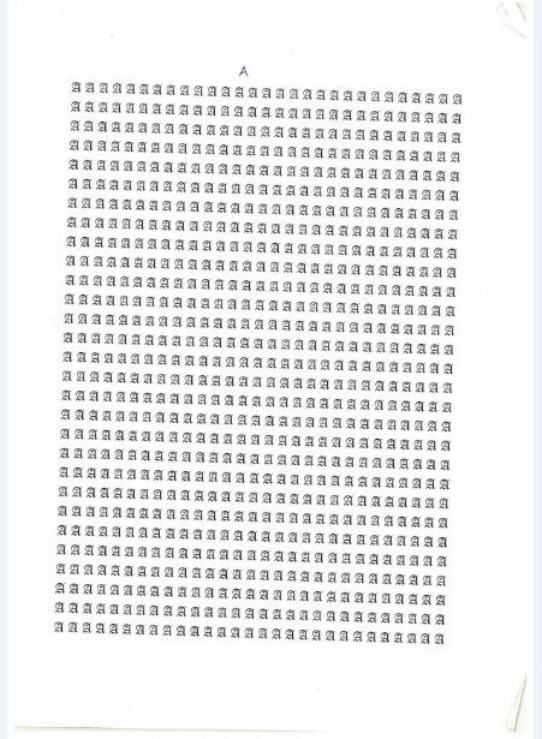

Fig -1: sample data set

3. PREPROCESSING

Pre-processing is done for the image which has been scanned. In the proposed system the pre-processing methods which have been used are binarization, noise removal, segmentation etc. Binarization will convert the gray scale images to binary images. It is basically used for foreground and background separation. Binarizarition will replace the image with black pixels, if the image intensity is less than the threshold value. If the image's intensity is greater than the threshold point, then the image will be replaced with white pixels. The image might contain noises even after the scanning is done; it is because of poor quality of scanner, printer, and old documents and so on. Noise removal is necessary for better recognition. A method is used to remove all the components which are lesser than P pixel from the binary image store as new binary image. Bounding Box methodology has been opted for the segmentation of the symbols. Technique finds the contours of the image and sequentially it creates a bounding box over a contour and segments it and saves the co-ordinates of the contour

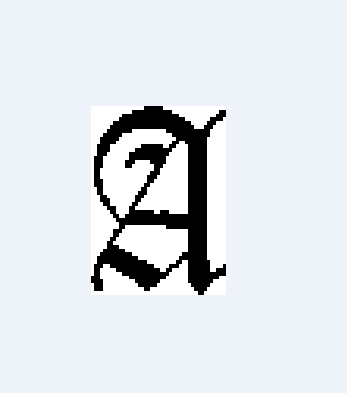

Fig -2: Sample data obtained after preprocessing

4. FEATURE EXTRACTION

The proposed algorithm Old English Font (OEF) is used for feature extraction and also very efficient. First the character is preprocessed and then the preprocessed character is put into a matrix which is a 5x4. The character image is resized into 50x64 using interpolation method. And the character image is divided into 16x16 zones. Identify the black pixels in each zone and store it. This procedure is done with all the similar images. So in total there will be 20 different features extracted.

5. CLASSIFICATION

To classify the characters we have used minimum distance classifier and closest feature vector is obtained. The Euclidean distance between the mean feature vector and the feature vector is settled and designates the numeral class to the nearest mean vector.

6. CONCLUSIONS

In this paper, we have explained the design and description of feature extraction algorithm for recognising Old English Font. The minimum distance classifier is used to recognise and classify the charaters. An overall accuracy rate of 79% is occurred with this technique. Future work will concentrate to achieve higher accuracy rate.

REFERENCES

[1] Ankit V.Patil and Mrinai M.Dhanvijay, 2015, 'Engraved Character Recognition Using ComputerVision To Recognize Engine And Chassis Numbers', International Conference on Information Processing (ICIP), 978-1-4673-7758-4/15/$31.00 ©2015 IEEE.

[2] ADITYARAJ, 2015, 'An Optical Character Recognition of Machine Printed Oriya Script',Third International Conference on Image Information Processing, 978-1-5090-01484/15/$31.00© 2015 IEEE.

[3] Shuwair Sardar and Abdul Wahab, 2010, 'Optical Character Recognition System for Urdu', 978-1-4244-8003-6/10/$26.00 ©2010 IEEE.

[4] D. N. HAKRO, A. Z. TALIB, Z. BHATTI and G. N. MOJAI, *2014* ,'A Study of Sindhi Related and Arabic Script Adapted languages Recognition', Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 46 (3) 323-334.

[5] Dr.Mrs.V.V.Patil, Rajharsh Vishnu Sanap and Rohini Babanrao Kharate, 2015, 'Optical Character Recognition Using Artificial Neural Network', International Journal of Engineering Research and General Science Volume 3, Issue 1, ISSN 2091-2730.

[6] Jack Greenhalgh and Majid Mirmehdi, 2015, 'Recognizing Text-Based Traffic Signs', IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 16, NO. 3.

[7] M. Abdul Rahiman and M. S. Rajasree, 2009, 'A Detailed Study and Analysis of OCR Research in South Indian Scripts' International Conference on Advances in Recent Technologies in Communication and Computing, 978-0-7695-3845-7/09 $25.00 © 2009 IEEE.

[8] VaidehiMurarka, Sneha Mehta, DishantUpadhyay, SonaliBhadra, and AbhijeetLal, 2012 ,' Recognition Using Pixel Probability Distribution', International Journal of Scientific & Engineering Research Volume 3, Issue 5, ISSN 2229-5518odd page.

[9] Chengshu (Eric) Li , 2015 ,'Handwritten English Alphabet Recognition Using Bigram Cost', International Journalof Computer Science and Information Technology IJCSIT 3.1 (2011): 27-38. Web.

[10] Prof. S.P.Kosbatwar Prof.S.K.Pathan, 2012,' Pattern Association for character recognition by Back-Propagation algorithm using Neural Network approach', International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.1.

[11] K.N. SARAVANAN, DR. R. ANITHA, 2014,Selective Subset of Relative Density Feature Extraction Algorithm for Unconstrained Single Connected

Handwritten Numeral Recognition, Australian Journal of Basic and Applied Sciences,Pages:429-435.

[12] Feng Pan, Mike Keane, 1994, A New Set of Moment Invariants for Handwritten Numeral Recognition, IEEE, CH-0-8186-6950, pp.154-158

[13] M.H. Shirali-Shahreza Karim Faez, Alireza Khotanzad, 1995, Recognition of Handwritten Persian/Arabic Numerals by Shadow Coding and an Edited Probabilistic Neural Network, IEEE, CH-0-8186-7310-9,pp.436-439

[14] Jianming Hu and Hong Yan, 1996, Structural Decomposition and Description of Printed and Handwritten Characters, Proceedings of ICPR,pp.230-234

[15] Thien M. Ha and Horst Bunke, 1997, Off-Line,Handwritten Numeral Recognition by Perturbation Method IEEE Transactions on pattern analysis and machine intelligence Vol. 19, No.5, pp. 535-539

[16] XUEFANG ZHU, C. GEOGER, 1998, An Off-line System for Recognition of Free Written Zipcodes,Proceedings of ICSP, pp.1253-1256

[17] Alceu de S. Britto Jr, Robert Sabouring, Flavio Bortolozzi and Ching Y. Suen, 2003,Complementary Features Combined in an HMMbased System to Recognize Handwritten Digits,Proceedings of the 12th International Conference on Image Analysis and Processing, ICIAP