International Journal of Scientific Research in Computer Science and Engineering
Vol.7, Issue.6, pp.40-42, December (2019)

Review Paper

E-ISSN: 2320-7639

Analysis of Text Recognition with Data Mining Techniques

Bindushree V.[1*], Rashmi G.R.[2], Uma H.R.[3]

[1,2,3]CS&E, BGSIT, ACU, Mandya, India

**Corresponding Author: bindushree11v@gmail.com, Tel.: 9964981481*

Available online at: www.isroset.org

Received: 12/Nov/2019, Accepted: 29/Nov/2019, Online: 31/Dec/2019

***Abstract*—** Recognization of text is a method that recognizes text from the file in the desired format (such as .doc or.txt).This process involves several steps, including pre-processing, segmentation, feature extraction, classification, and post-processing. The pre-processing is performed as a binarized image to convert a gray scale image, and noise is reduced on the input image of the basic operation performed by removing the noise of the image signal. The segmentation phase is used to segment the image given online and segment each character of the segmentation line. Feature extraction is to compute the characteristics of the image document. This document describes techniques for converting the textual content of a paper document into a machine-readable format. This paper analyzes and compares the technical challenges, methods, and performance of text detection and recognition studies in colour images.

***Keywords*—**Feature Extraction, Recognition

I. INTRODUCTION

Text recognition is important for many applications like automatic sign reading, navigation, language translation, license plate reading, content-based image search etc. So it is necessary to understand outlook text than ever. Texts in images carry high-level semantic information of the scene. Images in the webs and database are increasing. Developing effectual ways to handle and re-establish the ease of these resources is an urgent task. With the rapid growth of digital technology and devices manufactured by megapixel cameras and other devices such as Personal Digital Assistants (PDA), mobile phones, etc., are accountable for growing the consideration for information recovery and it leads to a new examine task. This paper gives us the survey on various methods used for text recognition, comparison study and also accuracy.

II. RELATED WORK

C. Patel and A. A. Desai [1] have proposed segmentation of text lines into words. They have used projection profile and morphological operations for segmentation. They have proposed zone identification for words. They have used distance transform method for identification of zone like upper, middle, and lower. They have proposed a handwritten character recognition system. They have used hybrid classifier using tree and k-NN. They have used structural and statistical features. They have achieved an accuracy of 63%.

A. A. Desai [4] has proposed character segmentation from old documents. He has used some pre-processing methods and Radon transform for segmentation. He has proposed a character recognition system for Gujarati numerals. He has used binarization, size normalization and thinning pre-processing methods. He has used hybrid features like a subdivision of skeletonized image and aspect ratio. He has used k-NN classifier with Euclidean distance method and achieved 96.99% accuracy. He has proposed similar work using profile vector-based features. He has used a multilayer feed forward neural network. He has achieved an accuracy of 82%.

M. Maloo and K. V. Kale [9] have proposed a handwritten numeral recognition system for Gujarati. They have used pre-processing methods like binarization, dilation, and skeletonization. They have used affine invariant moments (AMI) for feature extraction and SVM for classification and achieved 91% accuracy.

M. B. Mendapara and M. M. Goswami [10] have used binarization, noise removal, and thinning pre-processing methods. They have used stroke based directional feature and used k-NN as a classifier. They have achieved 88% accuracy.
R. Nagar and S. Mitra [11] have used binarization and thinning pre-processing methods. They have used orientation estimation features and SVM as a classifier and achieved 98.97% accuracy.

A. Vyas and M. Goswami [13] have used binarization, noise removal, and thinning pre-processing methods. They have

used modified chain code, Discrete Fourier Transform, and Discrete Cosine Transform as a feature. They have used k-NN, SVM and ANN as a classifier and achieved 85.67%,93.60%, and 93.00% accuracy respectively. Prutha Y M and Anuradha SG [14] have proposed a real-time traffic analysis system. They have used different morphological and edge detection techniques.

In Malayalam online handwritten character recognition, S.Joseph and A. Hameed [17] have used basic pre-processing methods and used six-time domain features with directional and curvature features. They have used SVM as a classifier and achieved 95.45% accuracy.

Anoop M. Namboodiri [18] has presented work on Malayalam and Telugu language. They have used normalization, resampling using a Gaussian low-pass filter and an equidistant resampling to remove variations in writing speed. They have used moments of the stroke, direction, curvature, length, an area of the stroke, aspect ratio as features. They have used SVM using a Decision Directed Acyclic Graph (DDAG) and discriminative classifier. They have achieved an accuracy of 95.78% on Malayalam and 95.12% on Telugu. Primekumar K.P. and S. Idiculla [19] have used duplicate point elimination, smoothing, normalization, resampling as pre-processing methods. They have used x-y coordinates; angular features, direction, and curvature are extracted. Using HMM classifier, they have used k means using Euclidean distance for training and using SVM classifier, they have used discrete wavelet transform for training. They have achieved an accuracy of 97.97% using SVM and 95.24% using HMM.

Fig 1: Comparison of different methods.

SL.No	Algorithm	Merits	Demerits
1.	Linear Regression	✓ Space Complexity is very low it just needs to save the weights at the end of training. Hence it is a high latency algorithm. ✓ It's very simple to understand ✓ Good interpretability ✓ Feature importance is generated at the time of model building.	✓ The algorithm assumes data is normally distributed in real but they are not. ✓ Before building multicollinearity should be avoided ✓ Prone to outliers
2.	Logistic Regression	✓ It is more robust. ✓ The independent variables don't have be normally distributed or have equal variance in each group ✓ It may handle non-liner effects	✓ It can't solve non-linear problems with logistic regression. Since its decision surface is linear ✓ Prone to over fitting
3.	Autoregressive integrated moving average	✓ Solid underlying theory stable estimation of time varying trends and seasonal patterns relatively few parameters.	✓ Non explicit seasonal indices hard to interpret coefficients, the danger of over fitting or misidentification if not used with care.
4.	Multivariate adaptive regression splines	✓ Works well even with large number of predictor variables. ✓ Automatically detects interaction between variables ✓ Efficient and fast robust to outliers	✓ Difficult to understand prone to over fitting ✓ Model is vulnerable to missing data

III. CONCLUSION

In this paper, an overview of various text recognition techniques, methods and recognition algorithms has been presented. Based on the literature review various text recognition algorithms accuracy are discussed. The detailed steps and flow of the text recognition techniques by surveying that image acquisition, pre-processing, feature extraction, classification, and post-processing from many research articles. Merits and demerits of text

REFERENCES

[1]. C. Patel and A. Desai, *"Segmentation of text lines into words for Gujarati handwritten text,"* Proc. 2010 Int. Conf. Signal Image Process. ICSIP 2010, pp. 130–134, **2010.**

[2]. C. Patel and A. Desai, *"Zone identification for Gujarati handwritten word,"* Proc. - 2nd Int. Conf. Emerg. Appl. Inf. Technol. EAIT 2011, pp. 194–197, **2011.**

[3]. C. Patel and A. Desai, *"Gujarati Handwritten Character Recognition Using Hybrid Method Based on Binary Tree-Classifier And K-Nearest Neighbour,"* Int. J. Eng. Res. Technol., vol. 2, no. 6, pp. 2337–2345, **2013.**

[4]. A. Desai, *"Segmentation of Characters from old Typewritten Documents using Radon Transform,"* Int. J. Comput. Appl., vol. 37, no. 9, pp. 10–15, **2012.**
[5]. A. A. Desai, *"Handwritten Gujarati Numeral Optical Character Recognition using Hybrid Feature Extraction Technique,"* Int. Conf. Image Process. Comput. Vision, Pattern Recognition, IPCV, **2010.**
[6]. A. A. Desai, *"Gujarati handwritten numeral optical character reorganization through neural network,"* J. Pattern Recognit., vol. 43, no. 7, pp. 2582–2589, **2010.**
[7]. A. a. Desai, *"Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space,"* CSI Trans. ICT, vol. 2, no. January, pp. 235–241, **2015.**
[8]. Mayil S. and Vanitha M, *"A Survey on privacy Preserving Data Mining Techniques"*, International Journal of Computer Science and Information Technologies. Vol.5 (5), pp. 6054-6056. ISSN: 0975- 9646, **2014.**
[9]. M. Maloo, K. V Kale, and I. Technology, *"Support Vector Machine Based Gujarati Numeral Recognition,"* Int. J. Comput. Sci. Eng.
[10]. M. B. Mendapara and M. M. Goswami, *"Stroke identification in Gujarati text using directional feature,"* Proceeding IEEE Int. Conf. Green Comput. Commun. Electr. Eng. ICGCCEE 2014, **2014.**
[11]. N. Rave and S. K. Mitra, *"Feature extraction based on stroke orientation estimation technique for handwritten numeral,"* in Eighth International Conference on Advances in Pattern Recognition (ICAPR), **2015.**
[12]. Manimaran R. and Vanitha M, *"An Efficient Study on Usage of Data Mining Techniques for Predicting Diabetes"*, International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol.3 (20), pp.268-272 ISSN: 2394-3785, **2016.**
[13]. A. N. Vyas and M. M. Goswami, *"Classification of handwritten Gujarati numerals,"* 2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015, pp. 1231–1237, **2015.**

AUTHOR'S PROFILE

Bindushree V, working as Assistant Professor at BGSIT. I have done Mtech in computer science and engineering. I am a member of "association of engineers group". I have published a paper on "image processing"

UMA H R, working as Assistant Professor at BGSIT. I have done Mtech in computer science and engineering

UMA H R, working as Assistant Professor at BGSIT. I have done Mtech in computer science and engineering