DOI: https://doi.org/10.53555/nncse.v2i3.501 Publication URL: http://nnpub.org/index.php/CSE/article/view/501

REVIEW PAPER ON VARIOUS METHODOLOGY OF TEXT EXTRACTION FROM IMAGE

Rucha Patil [1]*, Smruti Khati [2], Tulsi Thakur [3], Harshita Katragadda [4], Neha Ambulkar [5], Ketki Bhakare[6]
*[1,2,3,4,5]*Student, CSE, DBACER, Nagpur, India, [6]Assistant Professor, CSE, DBACER, Nagpur, India*
[1]6sspatil@gmail.com [2]khatismruti@gmail.com, [3]basiltulsi@gmail.com, [4]kharshita0294@gmail.com,[5]nehaambulkar@ymail.com, [6]ketki.bhakare@gmail.com

***Corresponding Author: -**
Email ID - *6sspatil@gmail.com*

Abstract: -

This review presents the various text extraction techniques and also compares the research results of various researchers in the domain of text extraction. A generic character recognition system has different stages like noise removal, skew detection and correction, segmentation, feature extraction and character recognition. Input is digitized image containing any text, which is preprocessed to segment it into normalized individual word and letters. The OCR, Neural Network, SVM are the various methods for text extraction. Text extraction helps to preserve history by making information efficiently searchable, easily manageable without the need for human labor.

Keywords: - *Feature Extraction, HMM, MDF, Neural Networks, OCR, Recognition, Segmentation, SVM.*

I. INTRODUCTION

Text extraction in images is an important research field in visual content understanding and retrieval, automatic explanation and structuring of images. A lot of work has been done for detecting text in images and a lot has to be done. A large number of techniques have been proposed to address this problem and the purpose of this paper is to classify and review various text extraction algorithms. Many existing paper-based collections, books, journals, etc. are converted to images. These images present many challenging research issues in text extraction and recognition.

Optical Character Recognition is a most important gift given by computer science to the mankind. It has made a lot of tedious work easy and speedy. OCR means a technique of recognition of machine printed or handwritten text by computer and then its conversion to an editable from as per the requirement. The goal of Optical Character Recognition (OCR) is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. The process of OCR involves several steps including segmentation, feature extraction, and classification.

II. LITERATURE REVIEW

Handwriting recognition is classified into two types as off-line and on-line recognition method. The work is carried out using LABVIEW. For recognition, pattern matching algorithm is used and segmentation method is used with neural network. The recognition rate of neural network is 99.90% for the handwritten word and for cursive handwriting with the accuracy of 7080%. Neural network recognizers learn from an initial image training set. The trained network then makes the character identifications. Each neural network uniquely learns the properties that differentiate training images. It then looks for similar properties in the target image to be identified. Neural networks are quick to setup; however, they can be inaccurate if they learn properties that are not important in the target data. [1]

Hidden Markov Model (HMM) for recognition of cursive handwritten English characters. By applying median filter, it avoids errors caused by noise in the scanned image. To reduce the complexities in the recognition process high quality samples are used. Here better results are obtained in terms of accuracy as well as speed. [2]

Muhammad Naeem Ayyaz1 [3]: Proposed a handwritten character recognition system based on a hybrid feature extraction technique has been presented. The system comprised three main stages, i.e. pre-processing, feature extraction technique, and SVM based training/classification. The proposed hybrid feature extraction technique, as experiments revealed, proved to capture local and global variations in handwritten character styles. The extracted feature vector was a combination of correlation function-based features and some statistical/structural features.

A new method is used in which 20,000 Hand written sample data files are used and 1200 for testing. The data is organized in class, separate class is decided for each character (i.e., vowels, consonants without modifiers, consonants with modifiers (%)). Then the result has been taken accuracy of each individual English Character is computed the table shows the average accuracy. In this paper the accuracy achieved is 98% for vowels, 97.50% for consonant without modifiers, 94% for consonants with modifiers. [4]

OCR is a technique used to convert any raster image of a document into a computer process able format. PCA (Principal Component Analysis) is used in the recognition phase which transforms a number of correlated features into a number of uncorrelated features. SVM is used for character recognition in classification phase. The maximum accuracy achieved by SVM classifier (RBF Kernel) with 10 fold cross validation technique is 91.95%. [5]

Feature extraction techniques generating local and global features are proposed by Vincirelli [6]. The local features are obtained from sub-images of the character including foreground pixel density information and directional information. The global features used included the fraction of the character appearing below the word baseline and the character’s width/height ratio.

Investigations of feature extraction techniques that may be applied to the classification of cursive characters for handwritten word recognition are presented in [7]. An MDF (Modified Directed Features) extraction technique was presented and was found to outperform the DF (Direction Features) extraction technique in terms of recognition accuracy. The first technique (DF) sought to simplify each character's boundary through identification of individual stroke or line segments in the image. The proposed MDF technique builds upon the DF technique described in Section A. The main difference is in the way the feature vector is created. For MDF, feature vector creation is based on the calculation of transition features from background to foreground pixels in the vertical and horizontal directions.

III. PROPOSED EXTRACTION SYSTEM

A. Training data

We will provide templates to the system so as to train it.

1. Image Acquisition

The input image is acquired by digital professional camera or scanned copy of an image.

2. Preprocessing

In this phase we perform noise removal, binarization, edge detection, dilation and filling so as to make it easy for OCR system to operate accurately. Prior to segmentation and recognition, it was necessary to preprocess all word images.

3. Feature Extraction

In this phase we extract a set of features, which maximizes the recognition rate with the least number of elements. The various techniques used for feature extraction are –

- *Moment*- Moments are statistical measure of the pixel distribution about the centre of gravity of the character.
- *Zoning*- Character matrix is divided into small portions or zones.
- *Projection Histogram*- It gives number of black pixels in the vertical and horizontal directions of the specified area of character.
- *N-tuples*- The position of black or white pixel in a character image is considered as a feature and provides number of important properties of pixel. After the completion of preprocessing & feature extraction we will create a separate database for each character.

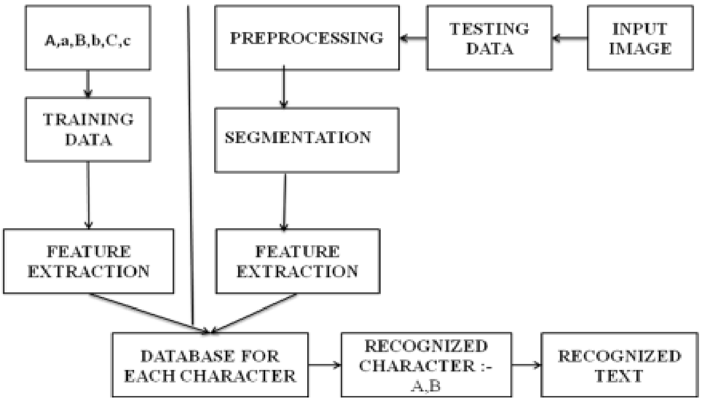

Fig.1: Work flow of extraction of text from image

B. Testing data

In this phase image is given as input & again preprocessing will be done. The output of preprocessing phase is provided as an input to segmentation phase.

1. Segmentation

Segmentation algorithm segments line by line and word by word detection. Prior to segmentation and recognition, it was necessary to pre-process all word images

2. Feature Extraction

The feature extraction will be done on the segmented characters, and then extracted features of segmented character will be match with the database already created in training phase.

3. Recognition

Once the features will match with the database the character will be successfully recognized and the output will be provided to the user.

IV. CONCLUSION

A large number of techniques have been proposed in the past but the detection of scene text with high precision and recall rate is still a challenging problem because of additional complexities such as varying lighting, variable font sizes, style, color, variance of orientation and complex background. The purpose of our paper is to review various techniques, discuss performance evaluation and to point out challenges for future research. The techniques like neural networks, structural and statistical pattern are available for recognition of text. But the major drawback of neural network is large training data which takes much more time to build this makes the neural network more complicated for a naive user to understand leading it to less user friendly. In this paper we are trying to presenting a way for extraction of text from image which is simple as well as user friendly. We are used the concept of matrix matching for recognizing the text.

References

[1].Pratibha A. Desai, Sumangala N Bhavikatti, Rajashekar Patil, "Neural Networks Based Offline Handwritten Character Recognition System", SARC-IRAJ InternationalConference,16thjune2013.

[2].Aparna.A , Prof.I.Muthumani," Optical Character Recognition for Handwritten Cursive English characters", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5, 2014.

[3].Muhammad Naeem Ayyaz, Imran Javed and Waqar Mahmood 1 2: Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction, Pak. J. Engg. & Appl. Sci. Vol. 10, Jan., 2012 (p. 57-67).

[4].Rohini B. Kharate *1, Dr.S.M.Jagade 2, Sushilkumar N. Holambe, "**A** Brief Review and Survey of Segmentation for Character Recognition", International Journal of Engineering Sciences, 2(1) January 2013.

[5].Shilpy Bansal, Mamta Garg, Munish Kumar, "A Technique for Offline Handwritten Character Recognition", IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014.

[6].F. Camastra and A. Vinciarelli, “Combining Neural Gas and Learning Vector Quantization for Cursive Character Recognition”, Neurocomputing, vol. 51, 2003, pp. 147-159.

[7].M. Blumenstein and X. Y. Liu, B. Verma, “A Modified Direction Feature for Cursive Character Recognition “.

.