International Journal of Computer Science and Engineering Communications,
Volume.5, Issue.5 (2017): Page.1709-1714
www.ijcsec.com

Improved Handwritten Text Feature Extraction using HOG and SVM Feature

Femin.P.D[1], Krishnapriya.K.S[2] and Vince Paul[3]
[1]PG Scholar, [2]Assistant Professor [3],Professor
[1,2]Department of Computer Science and Engineering,
[1,2]Sahrdaya College of Engineering & Technology, Kodakara,Kerala, India
femin.pottekkatt@gmail.com[1], krishnapriyaks@sahrdaya.ac.in[2], vinceakkara@gmail.com[3]

Abstract: *Handwritten Text Extraction (HTE) is an emerging technology in image processing. HTE is the wonders of empowering a machine to automatically realize the characters written in a user dialect. Optical character extraction has turned out to be a standout amongst the best candidates of innovation in the field of pattern recognition and artificial intelligence. In this paper we propose a methodology for extracting and recognizing characters or text using Histogram of Gradient (HOG) feature extraction and Support Vector Machines (SVM) as classifier. To evaluate the accuracy of our proposed method we use the ground truth data.*

Keywords *- Handwritten Text Extraction, Histogram of Gradient, Support Vector Machines.*

I. Introduction

Text recognition is a territory of example distinguishing proof that has been the subject of extensive amid the current decades. Manually written text shows wide complex varieties. Penmanship is a standout amongst the most important mean in day by day discussion. Amid the current years, by far most of the conspicuous field of study and applications joined for bank check taking care of, sent wraps address perusing, and composed by hand message recognizable proof in records and recordings [1]. Character Recognition System (CRS) used to distinguish mortal print image. These images might be alphabetic, numeric, punctuation and so forth. These images might be either printed or composed by submit an assortment of various size and textual style. All the more exactly character recognition is the way toward distinguishing and perceiving character from information picture and change it into American Standard Code for Information Interchange or other comparing machine editable form [2]. The chore of recognition comprehensively segregates into two kinds: written by hand and machine printed. The printed character reference is uniform and extraordinary. Manually written alphabets are not uniform and the size and shape may depend upon the pen utilized by the essayist. Handwriting of same corporal also may vary be based upon the billet on which the mortal composite. Different written work styles prompt the bending in example from the standard examples used to equipping the game plan, giving fake outcomes. So it is a troublesome undertaking to outline framework, which is equipped for perceive the character with higher precision.

CRS is classified into two different types:

- Off-line CRS
- On-line CRS

On-line CRS recognize by perceiving penmanship recorded with a digitizer as a period succession of pen directions[3]. In the event of online CRS character identification, the penmanship is caught and put away in advanced shape by means of various ways. That sort of information is known as digital ink and can be viewed as a dynamic portrayal of handwriting. The got signal is changed over into letter codes which are usable inside PC and text handling applications. Disconnected penmanship identification suggest to the way toward perceiving words that have been checked from a surface and are put away carefully in dark scale administrate. Off-line CRS recognize from the scanned images. This will use feature extraction methods and classifiers to extract and realize the characters. What's more, as of today, OCR engines are fundamentally considered machine printed and ICR for hand printed. There is no OCR engine that supports penmanship realization as of today. The CRS can be partitioned into recognition of manually written and printed. HTR is harder to actualize than printed character due to variety human handwriting styles and customs. In printed character realization, the images handled are in the category of standard fonts.

The CRS is divided into four stages.

1. Pre-processing
2. Text Localization
3. Feature Extraction
4. Recognition.

The purpose of pre-processing is to discard irrelevant information in the input data, that can negatively affect the recognition. This concerns speed and accuracy. Pre-processing usually consists of binarization, normalization, sampling, smoothing and denoising. Text localization is done using region extractor. Next Features are extracted from the localized regions. Then apply classification algorithms to recognize the input. Paper is organized as follows. Section II describes related work of CRS using different operations. Section III describes the proposed methodology. Section IV presents experimental results showing results of images tested. Finally, Section V presents conclusion.

II. RELATED WORK

Analysts have been contemplating on the acknowledgment of manually written records since 1960s. A review on the written by hand record acknowledgment by R. Plamondon and S. N. Srihari began in mid-2000. Look into papers on this subject have been distributed by the scientists since 1960s and the most recent one in 2015. In the 2011 Anshal[4],Plamondon[5],Arica[7] concentrates particularly on disconnected acknowledgment of manually written English words by first distinguishing singular characters. They considered two methodologies for the word acknowledgment. It was all encompassing methodology and division based approach. In the 2012 Neeta[6] a Diagonal based element extraction procedure was utilized alongside the neural systems for the acknowledgment of manually written archives. A bolster forward counterfeit neural system is being utilized for character grouping.

In 2013 Kandula[8], Takumi[9]two systems were utilized to recognize the transcribed character and they were the Active Character Detection and the Contour Algorithms. Once the form of a given example is removed, it's diverse attributes will be inspected and utilized as components which will later on be utilized as a part of example characterization. The 2014 Disha[10], Sandeep[11],Prerna[12], depended on acknowledgment of transcribed characters within the sight of commotion. With a specific end goal to defeat this confinement, the back spread of manufactured neural system is intended for the English character acknowledgment in nearness of clamor.

III. Proposed Methodology

The proposed CRS has two phases: Training phase and Testing phase.

Workflow of Training phase

1. Take the 25 images of each character
2. Pre-process using Gaussian blur
3. Extract the HOG feature of each image
4. Make data file from feature vector and its label
5. Create a model file from the data file using svm-train

Workflow of Testing phase

1. Take the input image
2. Pre-process using Gaussian blur
3. Perform horizontal projection profile
4. Perform vertical projection profile
5. Extract the HOG feature of each character
6. Predict this character using svm-predict against the model file created during training

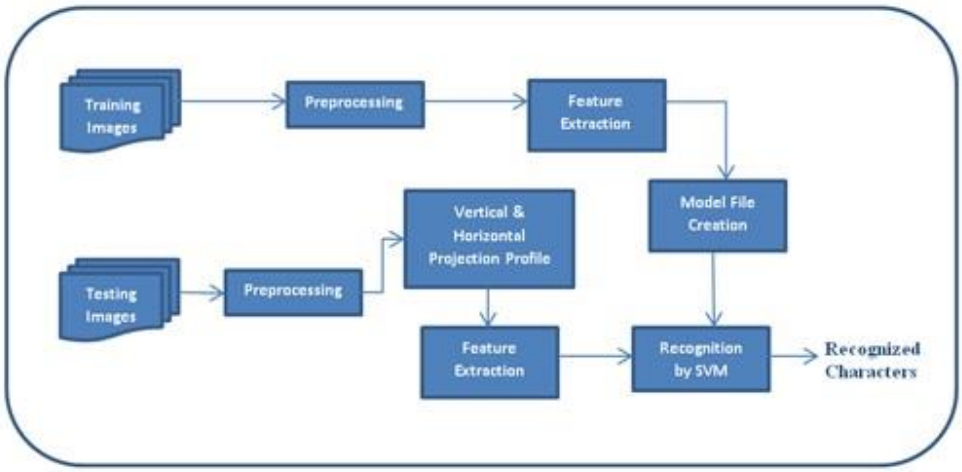

Figure 1 Proposed block diagram

Block diagram of proposed work is shown in Figure1. It contains training and testing phases. Model file created during training is used for prediction in the testing phase and identify the characters.

First perform pre-processing and text localization. After this, connected component identification performed using projection profile. Next extract features and test against the model file created during offline training phase. Predicted characters will be print in a text file.

IV. Experimental results

Dataset: The Chars74K dataset used in this for training. 25 images of each alphabet used for training. Experimental Setup: The complete CRS implemented using OpenCV with C++. Visual studio is used as C++ IDE. This IDE provides user friendly environment for coding, testing and debugging. OpenCV(OpenSourceComputerVision) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel's research center in Nizhny Novgorod (Russia), it was later supported by Willow Garage and is now maintained by Itseez. It is cross-platform and free for use. Figure2 shows the input image and its corresponding bounded character image. For plotting the boundary of each character, contour detection is used. Input image is the scanned image of handwritten document.

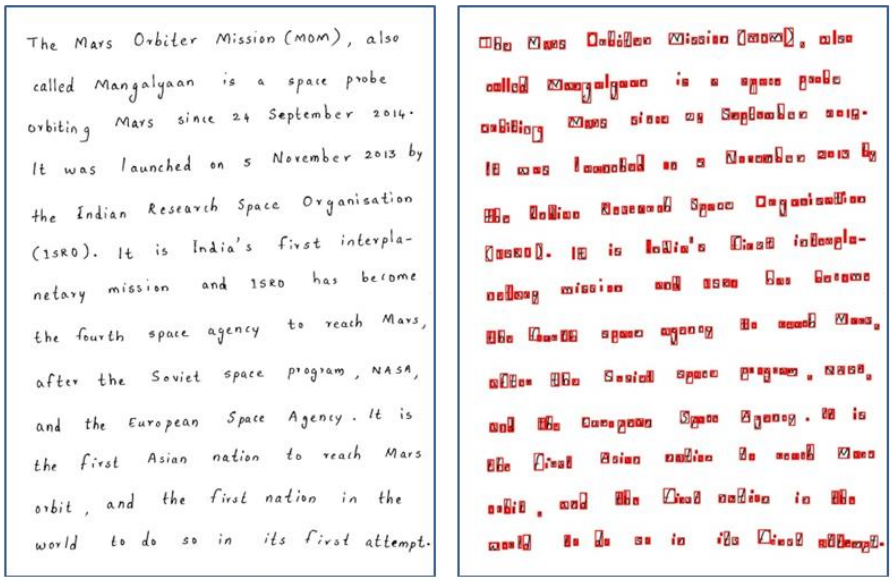

Figure 2 Input image and its corresponding bounded rectangles

Result: The complete images in the dataset used for training. In the testing phase each character of input image is predicted using the model file created during training. Accuracy is calculated by counting the input characters and the correct output characters. Figure 3 shows the graph of each image with its corresponding accuracy. Accuracy is determined by taking ratio of correct characters identified to the total number of characters.

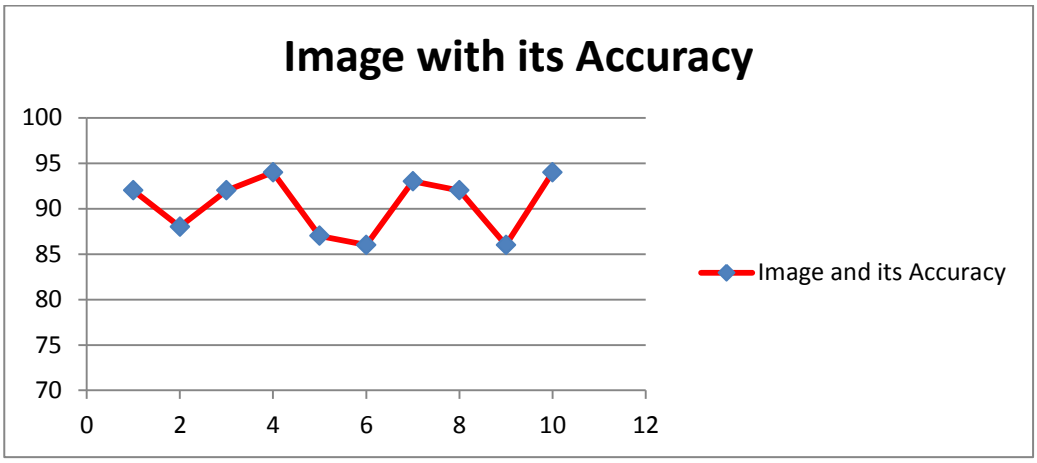

Figure 3 Performance Analysis

V. Conclusion

We have implemented character extraction system which recognizes all characters in the input image. Our algorithm localizes the text region from the image and identifies or recognizes each character using SVM classifier.

REFERENCES

[1]Siyang Qin and Roberto Manduchi , "A Fast and Robust Text Spotter",IEEE Winter Conference on Applications of Computer Vision (WACV),2016, pp.1-8.
[2]Jay H. Bosamiya, PalashAgrawal, ParthaPratim Roy and R. Balasubramanian , "Script Independent Scene Text Segmentation using Fast Stroke Width Transform and GrabCut", 3rd IAPR Asian Conference on Pattern Recognition (ACPR),2015, pp.151-155.
[3]Femin P D and Dr. Vince Paul, "A Comparative Study of Techniques Used in Handwritten Character Recognition", International Journal of Innovative Research in Science, Engineering and Technology Vol. 5, Issue 11,November 2016, pp.20040-20045.
[4]Anshul Gupta,Manisha Srivastava and Chitralekha Mahanta, "Offline Handwritten Character Recognition Using Neural Network"AI 2011 International Conference on computer applications and industrial electronics, April 2011,pp.102-107.
[5]Plamondone ,S.Chaturved, N.R.Sondhiya, R.N.Titre and Izhikevich, "Model Based Pattern Classifier for Hand Written Character Recognition "– A Review Analysis, International Conference on Electronic Systems, Signal Processing and Computing Technologies (ICESC),Jan. 2014, pp.346-349, 9-11.
[6]Neeta Nain and Subhash Panwar, "Handwritten Text Recognition System Based onNeural Network",Journal of computer and information technology , vol2,no-2,June 2012, pp.95-103.
[7]Arica, N., and Yarman-Vural. F.T., "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Trans. On Systems, Man, and Cybernetics, Vol. 31. No. 2,2001, pp. 216-233.

[8]Kandula Venkata Reddy , D. Rajeswara Rao , K. Rajesh "Hand Written Character Detection by Using Fuzzy Logic Techniques",International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3,March 2013, pp.256-260.

[9]Takumi Kobayashi, "BoF meets HOG: Feature Extraction based on Histograms of Oriented p.d.f Gradients for Image Classification", IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp.747-754.

[10]Disha Bhattacharjee , Deepti Tripathi, Rubi Debnath, Vivek Hanumante, Sahadev Roy "A Novel Approach for Character Recognition", International Journal of Engineering Trends and Technology (IJETT) – Volume 10 Number 6,Apr 2014, pp-271-275.

[11]Amrita hirwanil, Sandeep Gonnade, "Handwritten Character Recognition System Using Neural Network",International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 2,February 2014 , pp.83-88.

[12]Prerna Kakkar, Umesh Dutta "A Novel Approach to Recognition of English Characters Using Artificial Neural Network", International Journal of Advanced Research in Electrical, Electronics, Volume-3, Issue-9,Sep-2015, pp.28-33.