ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)
Vol. 2, Issue 2 (Apr. - June 2015)
International Journal of Advanced Research in Education Technology (IJARET)

An Improvement Study for Optical Character Recognition by using Inverse –SVM in Image Processing Technique

[I]Dinesh KumarVerma, [II]Anjali Khatri
[I]Assistant Professor (ECE) PDM College of Engineering, Bahadurgarh, Haryana, India
[II]M.Tech student (ECE) PDM College of Engineering, Bahadurgarh, Haryana, India

Abstract
In this paper, we consider the new method for isolated handwritten characters and numerals recognition using INVERSE-SVM. The Inverse support vector machine support vector machine which is based on statistical learning theory, with good generalization ability is used as the classifier. This method is robust to scale and frame size changes. Experimental analysis is conducted on a database obtained by combining the database with alphanumeric where the I-SVM combination. A morphological procedure follows on the OCR alphanumeric using sobel filtering then we need canny edge detection. After that we used the inverse SVM such as erosion, dilation, closing & opening for text extraction and OCR technique for character recognition from image. The results indicate that the N-SVM system gives the best performance in terms of training time and error rate.

Keyword
OCR, I-SVM, Sobel, Canny Edge Detection etc.

I. Introduction

The history of OCR research is relatively old in the era of pattern recognition. Optical Character Recognition has earned interest as it renders services for various systems. Some practical applications of OCR include recognition of vehicle number plates for identification and security concern information, in banks for processing checks, for easy-search of the scanned documents in database, finds implementation in hand-writing
Recognition and use in T2MSTA[1]. This technology has been exploited in almost various areas in industries like robot vision, Automatic bank check processing [2]. Now a days it use in automated analysis of medical image of blood contents, identification of humans from figure prints etc [3].

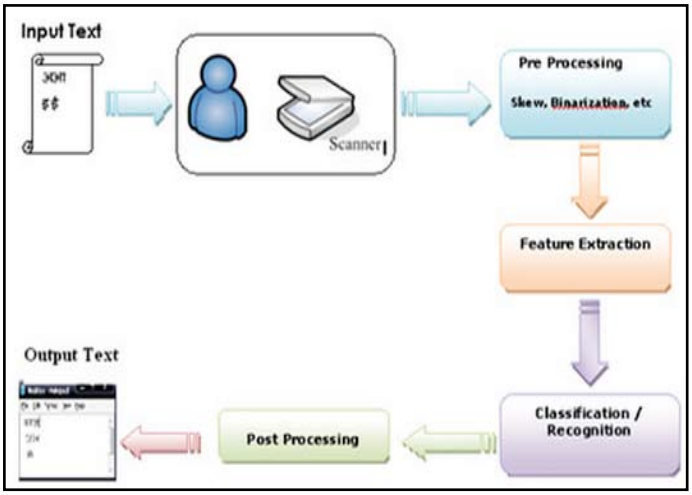

Fig.1: A flow chart of OCR

Optical Character Recognition (OCR) is a field of research in pattern recognition, artificial intelligence and machine vision, signal processing. Optical character recognition (OCR) is usually referred to as an off-line character recognition process to mean that the system scans and recognizes static images of the characters. It refers to the mechanical or electronic translation of images of handwritten character or printed text into machine code without any variation.

A. Phases of General Character Recognition System

Binarization

Document image binarization is usually performed in the preprocessing stage of different document image processing related applications such as optical character recognition (OCR) and document image retrieval. It converts a gray-scale document image into a binary document image and accordingly facilitates the ensuing tasks such as document skew estimation
and document layout analysis. As more and more text documents are scanned, fast and accurate document image binarization is becoming increasingly important.

B. Pre-processing (Salt paper)

To obtain an image with 'speckle' or 'salt and pepper' noise we need to add white and black pixels randomly in the image matrix. First convert the RGB image into gray scale image. Then generate random values for the size of the matrix. Here I used MATLAB function 'ran dint'. IN The pre-processing phase, there is a series of operations performed on the scanned input image. It enhances the image rendering it suitable for segmentation the gray-level character image is normalized into a window sized. After noise reduction, we produced a bitmap image. Then, the bitmap image was transformed into a thinned image.

Segmentation

The Segmentation phase is the most important process. Segmentation is done by separation from the individual characters of an image. Segmentation of handwritten characters into different zones (upper, middle and lower zone) and characters is more difficult than that of printed documents that are in standard form. This is mainly because of variability in paragraph, words of line and characters of a word, skew, slant, size and curved. Sometimes components of two adjacent characters may be touched or overlapped and this situation creates difficulties in the segmentation task.

C. Feature Extraction

In this phase, features of individual character are extracted. The performance of an each character recognition system that depends on the features that are extracted. The extracted features from input character should allow classification of a character in a unique way. We used diagonal features, intersection and open end

ISSN : 2394-2975 (Online)
ISSN : 2394-6814 (Print)

points features, transition features, zoning features, directional features, parabola curve fitting–based features, and power curve fitting–based features in order to find the feature set for a given character.

II. Litrature Review

A detailed survey on research work on Indian languages is presented in [3-4].In this paper, properties of Indian scripts, methods and approaches applied to recognize characters are discussed. Vikas **Dungre et al. [5]** reviewed feature extraction using Global transformation and series expansion like Fourier transform, Gabor transform, moments; statistical features like zoning, projections crossings and distances; and some geometrical and topological features commonly practiced. PrachiMukherji and **PritiRege [6]** have used structural features like endpoint, cross-point, junction points, and thinning. They classified the segmented shapes or strokes as left curve, right curve,horizontal stroke, vertical stroke, slanted lines etc.

GiorgosVamvakas et al. [7], [8] described the statistical and structural features they have used in their approach of Greek handwritten character recognition. The statistical features they have used are zoning, projections and profiling, and crossings and distances. By zoning they derived local features and also described in- and out- profile of contour of images. The structural features they depicted are end point, crossing point, loop, horizontal and vertical projection histograms, radial histogram, out-in and in-out histogram.

Sarbajit Pal et al. [9] have described projection based statistical approach for handwritten character recognition. They proposed four sided projections of characters and projections were smoothed by polygon approximation.

Nozomu Araki et al. [10] proposed a statistical approach for character recognition using Bayesian filter. They reported good recognition performance in spite of simplicity of Bayesian algorithm

III. Problem Statement

In this paper we face the some basic problem during the implement:

Input Image: These are main sources of noise in the input image.

- Noise in dataset image.
- Text that is not digital is virtually invisible
- OCR (optical character recognition) technology does not produce satisfactory results for historic documents.
- There is a lack of institutional knowledge and expertise which causes "re-inventing the wheel "Innovate OCR software and language technology.
- Optical Character Recognition deals with the problem of recognizing optically processed characters.
- Problems in thresholding: Top: Original greylevel image, Middle: Image thresholded with global method, Bottom: Image thresholded with an adaptive method.

IV. System Model

- Numerical optimization in a high-dimensional space suffers from the curse of dimensionality. This computational problem is avoided by using the notion of an inner-product kernel (defined in accordance with Mercer's theorem) and solving the dual form of the constrained optimization problem formulated in the input (data) space.

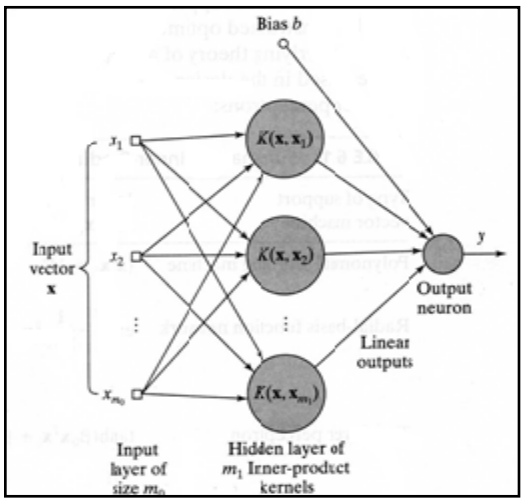

Fig. 2 : architecture of SVM

Consider training sample

Where is the input pattern, is the desired output:

$$\begin{cases} W_0^T X_i + b_0 \geq +1 & for \quad d_i = +1 \\ W_0^T X_i + b_0 \leq -1 & for \quad d_i = -1 \end{cases}$$

It is the equation of a decision surface in the form of a hyper-plane.

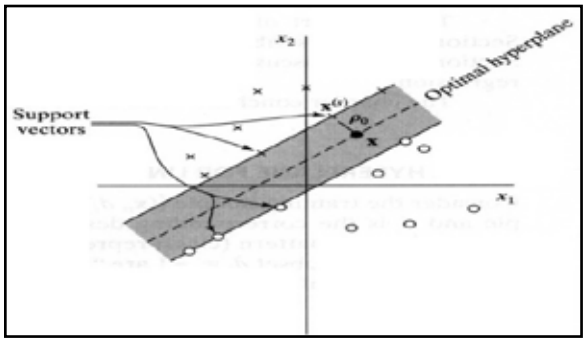

Fig. 3: The closest data point

- The closest data point is called the margin of separation
- The goal of a SVM is to find the particular hyper-plane of which the margin is maximized
- Optimal hyper-plane
- $W_0^T X + b_0 = 0$
- **Given the training set** $T = \{(\vec{x}_i, d_i)\}$
- **The pair must satisfy the constraint:**

$$\begin{cases} \vec{W}_0^T \vec{x}_i + b_0 \geq +1 & for \quad d_i = +1 \\ \vec{W}_0^T \vec{x}_i + b_0 \leq -1 & for \quad d_i = -1 \end{cases}$$

- The particular data point for which the first or second line of the above equation is a satisfied with the equality sign are called support vectors

V. Proposed Implementation

Before running the OCR we have to perform some basic steps for image processing. For improves the OCR efficiency and Performance.

A. Edge Detection

The first step in processing the image from the canny is the edge detection.

B. Dataset preparation & preprocessing

Samples of each alphanumeric consonant have been written by each people means each people have written 360 (10*36) alphanumeric characters & numerals in A4 size sheet. After that this sheet is scanned and saved as jpeg image (gray-scale image).

C. Following steps have been performed in order to pre - process the image before feature extraction:-

1. Intensity values of an image were adjusted.
2. Images were converted into binary images by choosing threshold value 0.5.
3. All connected components (objects) that have fewer than 30 pixels were removed from the binary images.
4. Sobel filtering, which is a nonlinear operation often used in image processing to reduce "salt and pepper" noise was applied on all images.
5. Each Image was segmented horizontally by finding the black pixel in each row.
6. After horizontal segmentation, each line was segmented vertically and we obtained our required alphanumeric image.
7. Finally all images of alphanumeric were normalized to size 90*90.
8. Preprocessing using Salt paper apply.

D. Feature extraction techniques

Feature extraction is an integral part of any recognition system. The aim of feature extraction is to describe the pattern by means of minimum number of features that are effective in discriminating pattern classes.

The sobel operator is very similar to Prewitt operator. It is also a derivate mask and is used for edge detection. Like Prewitt operator sobel operator is also used to detect two kinds of edges in an image:

Vertical direction

Horizontal direction

E. Classifier (SVM)

Support Vector Machine is supervised Machine Learning technique. It is primarily a two class classifier. Width of the margin between the classes is the optimization criterion, i.e. the empty area around the decision boundary defined by the distance to the nearest training pattern.These patterns called support vectors, finally define the classification function. All the experiments are done on LIBSVM 3.0.1[20]

which is multiclass SVM and select RBF (Radial Basis Function) kernel. A feature vector set fv(xi) i=1…m, where m is the total number of character in training set and a class set cs(yj) j=1…m , cs(yj) { 0 1 ….9} which defines the class of the training set, fed to Multi Class SVM.

F. OCR

Finally, the prepared image is fed to an open-source OCR method, Tesseract [10], which processes the image and reads the text on it [1]. The text is then compared to the expected text, e.g. expected menu content [1]. If the extracted text matches the expected one, the test passes, on the contrary it fails [1].

VI. Result

In our result initially we take input image which configure by input text and fetch which are as in alphabetical and numerical formation, that all results are below:

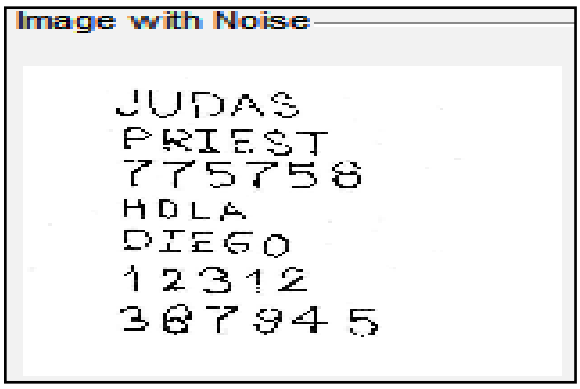

Fig. 1: Input Image with Noise

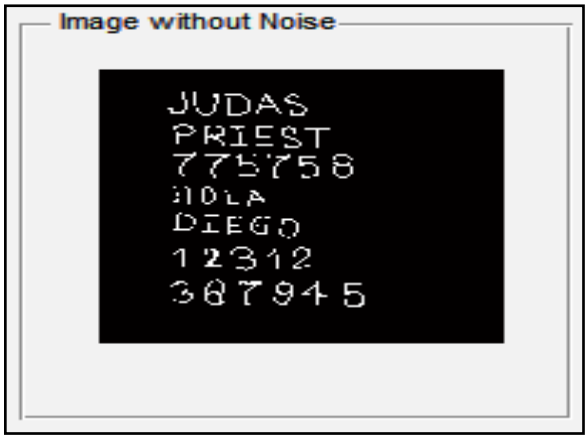

Fig. 2: Image without Noise by using Gaussian

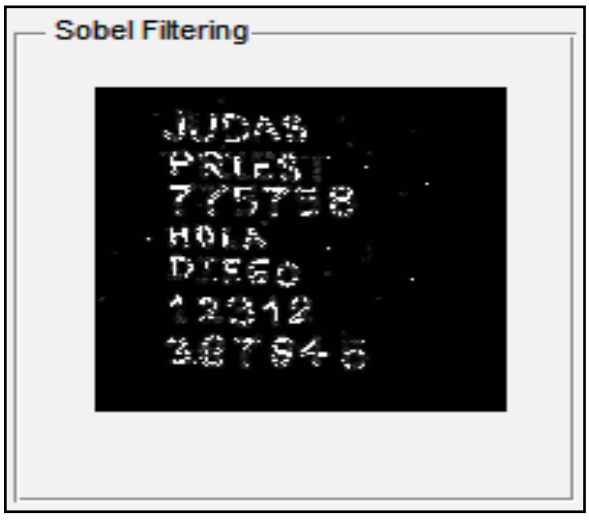

Fig. 3: apply sobel filter text

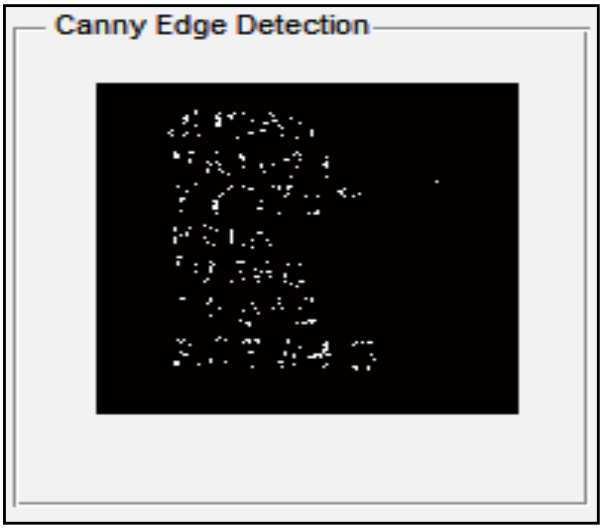

Fig. 4: canny edge detection for image text

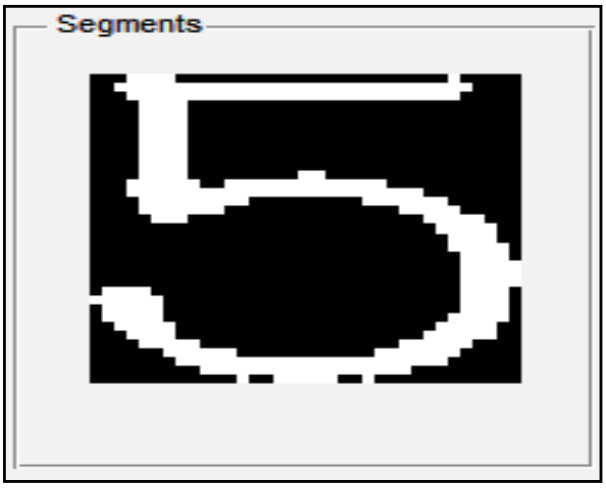

Fig. 5: Segmentation of text for image text

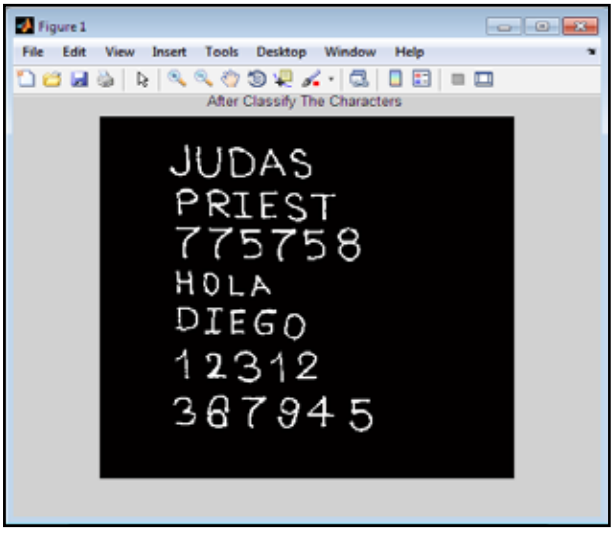

Fig. 6: Classify by I-SVM for text

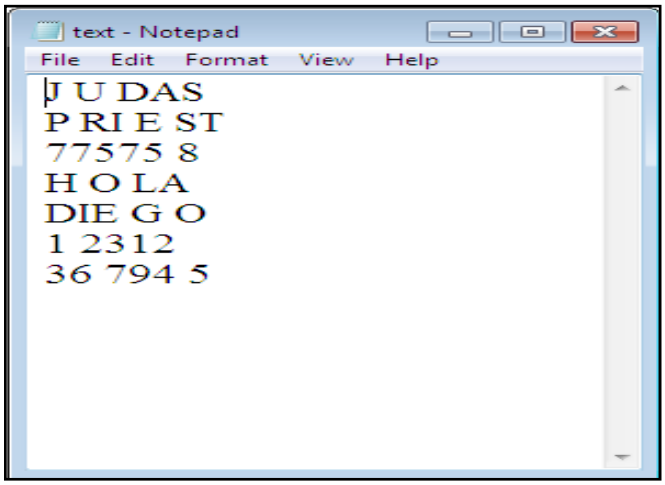

Fig. 7: Text extract in to notepad window

VII. Conclusion

This paper tells about OCR system for handwritten character & numeric recognition. Preprocessing techniques used in document images as an initial step in character recognition systems were presented. Thefeature extraction step of optical character recognition is the most important. It can be used with existing OCR methods, especially for English text.

Future work

The so far developed system could become even more efficient and reliable by eliminating some problems for further enhancement

1. Detection of white-space between the words.
2. Some lower cases are not recognized correctly like 'i'. It is recognized as a combination of more than one letter.
3. The system get confused while recognizing some similar looking lower and upper cases like 'o' 'u' 'c' and some similar looking alphabets and numbers like '2' and 'z', '9' and 'q'.
4. Also we would like to further recognize the numeric no i.e 1/2 …….1/9 etc

References

[1] *VivekHanumante, RubiDebnath, DishaBhattacharjee, DeeptiTripathi and Sahadev Roy "English Text to Multilingual Speech Translator Using Android," International Journal of Inventive Engineering and Sciences, vol-2, no-5, pp 4-9, April 2014.*

[2] *R. F. P. Neves, C. A. B. Mello, M. S. Silvae and B. L. D. Bezerra, "Thresholding the Courtesy Amount of Brazilian Bank Checks Based on Tsallis Entropy," Latin America Transactions, IEEE, (Revista IEEE America Latina) , vol.7, no.6, pp.726-731, Dec. 2009.*

[3] *Onal Scripts: A Survey of Offline Techniques," ACM Transactions on Asian Language Information Processing, Vol. 11, No. 1, Article 1, Publication date: March 2012.*

[4] *Vikas J Dungre et al., "A Review of Research on Devnagari Character Recognition", International Journal of Computer Applications (0975-8887), Volume-12, No.2, November 2010.*

[5] *PrachiMukherji, PritiRege, "Shape Feature and Fuzzy Logic Based Offline Devnagari HandwrittenOptical Character Recognition", Journal of Pattern Recognition Research 4 (2009) 52-68, 2009.*

[6] *Vamvakas, G.; Gatos, B.; Petridis, S.; Stamatopoulos, N.; "An Efficient Feature Extraction and Dimensionality*

Reduction Scheme for Isolated Greek Handwritten Character Recognition," Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, vol.2, no., pp.1073-1077, 23-26 Sept. 2007.

[7] Vamvakas, G.; Gatos, B.; Petridis, S.; Stamatopoulos, N.; et al., "Optical Character Recognition for Handwritten Characters" ppt, [Online].Available:http://www.iit.demokritos.gr/IIT_SS/Presentations/OffLine%20Handwritten%20OCR.ppt. Access ed in 2010.

[8] Sarbajit Pal, JhimliMitra, SoumyaGhose, Paromita Banerjee, "A Projection Based Statistical Approach for Handwritten Character Recognition," in Proceedings of International Conference on Computational Intelligence and Multimedia Applications, vol. 2, pp.404- 408, 2007.

[9] Araki, N.; Okuzaki, M.; Konishi, Y.; Ishigaki, H.;, "A Statistical Approach for Handwritten Character Recognition Using Bayesian Filter," Innovative Computing Information and Control, 2008. ICICIC '08. 3rd International Conference on , vol., no., pp.194, 18-20 June 2008.