Optical Character Recognition

YashPandey[1],BhanuPratap[2],Sangras Bhargav[3] ,J.Shiva Nandhini[4]

B.Tech CSE-STUDENTS , 3rdyr , SRMIST ,RAMAPURAM,CHENNAI.[1][2][3]

Asst.Professor , CSE , SRMIST ,RAMAPURAM,CHENNAI.[4]

Abstract

Handwriting Recognition has always played a prominant role in active and challenging area of research & development. Handwriting recognition system plays a significant part in today's modern technological enhanced world. Handwriting recognition is very popular and has expensive computational executive work. In present time, it is very hard to find actual meaning of handwritten documents. There are plenty areas where it needs to recognise the correct words, alphabets and digits. Handwriting recognition has two basic type existing one is online and other is offline.

In this project, by using Linear Support Vector we will present the handwriting recognition system in a very simple and feasible way. Different methods are studied and included it to enhance ease usability of programmed evaluation framework and improvise it with the help of SVM , Machine learning & Artificial intelligence & thus we enhance the character recognition system.

Keywords— *SVM , HCR , OCR , k-nn model , Markov model*

I. INTRODUCTION

Optical Character Recognition (OCR) which involves in photoscanning of a character-by-character,the analysis of a scanned-in image, and then the translation of the character in photos into character codes, such as the ASCII, commonly used in data processing . It also involves automatic conversion of the handwritten text which is an image into the letter codes which are usable within the computer and its text-processing of applications .It has the three basic steps: Detecting & Recognising the input , error/disruption removal & processing of digital output .

As sum provides the number of different ways to use of all the features of every existing systems in our field to gather the system which is of much more efficient and very easy to use . Generally , an OCR system which comprises of either the recognising as handwritten data such as the scripts or fonts or like any printed image .

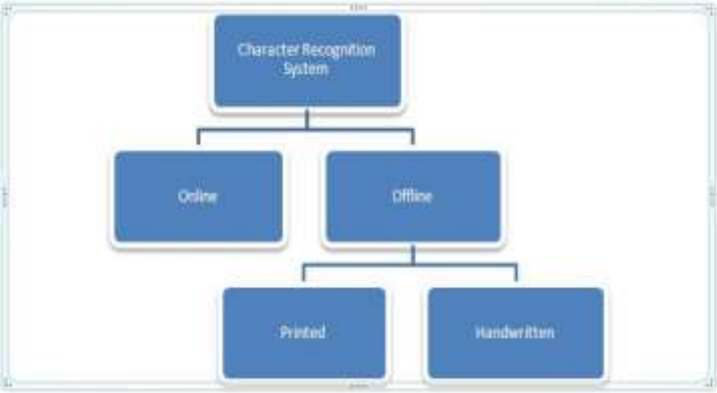


fig.1.1 OCR System

Offline character recognition methods discussed in two ways as given by fig1.1:.

(A)Printed Character Recognition:

It involves extracting data from scanned documents, camera images or image only PDFs, enabling you to access and edit the content . This system is enormous becausethey enable usersto harness the power of computers to access printed documents.

(B)Handwritten Character Recognition:

It involves extracting data from handwritten scripts and enable user to access and edit the content .

II. RELATED WORK

We develop an optical character recognition (OCR) engine for handwritten Sanskrit using a two-stage classifier. Inside the standard OCR pipeline, we focus on the classification problem assuming characters have been preprocessed decently.One challenge we face is that the language of Sanskrit has about a hundred core characterswhere model driven methods, like Support Vector Machine (SVM), have to search in the exponentially growth of the combinatoric model space during training, while data driven methods, like k nearest neighbor (kNN), becomes costly in computation during testing.[1]

We propose a method of feature extraction for digit recognition that is inspired by vision research: a sparse-coding strategyand a local maximum operation . We show that our method, despite its simplicity, yields state-of-the-art classification results on a highly competitive digit-recognition benchmark.We first employ the unsupervised Sparsenet algorithm to learn a basis for representing patches of handwritten digit images. We then use this basis to extract local coefficients. In a second step , we apply a local maximum operation in order to implement local shift invariance. Finally, we train a Support - Vector - Machine on the resulting feature vectors and obtain state-of-the-art classification performance in the digit recognition task defined by the MNIST benchmark.[3]

Numerous progressions have occurred in the area of Offline Handwriting Recognition, for example, Feature Extraction Techniques, Character Recognition Techniques and so on. HCR, Handwriting Character Recognition is the ability of a framework to interpret intelligible handwritten input from sources, for example, paper records, photos and might be sensed offline by Optical Scanning and Intelligent Word Recognition. Likewise OCR, Optical Character Recognition is the mechanical or electronic transformation of pictures of typed, manually written or printed content into machine-encoded content. It is a typical technique for digitizing printed writings and utilized as a part of machine procedures, for example, Cognitive Computing, Machine Translation and so forth. It is a field of research in Pattern Recognition, Artificial Intelligence and Computer Vision.[4]

Character recognition (CR) has been widely studied in the previous half century and develops to a level sufficient to produce technology determined applications. Nowadays, the precise recognition of machine printed characters is considered mainly a solved trouble. A lot of commercial products are paying attention towards that way, achieving high recognition rates. At the recognition phase a features are extracted from Characters in order to classify them to predefined classes techniques. However, handwritten character recognition is reasonably difficult. So, handwritten recognition documents is still a subject of active researchNow, the quickly growing -computational power enables the execution of the present CR methodologies and creates a rising demand on many promising application field, which require more highly developed methodologies.[2]

Commercial OCR tool Transym OCR by considering vehicle number plate as input. From vehicle number plate we tried to extract vehicle number by using Tesseract and Transym and compared these tools based on various parameters. [1] .Gabor wavelets and principle component analysis. We conclude that the learning ofa sparse representation oflocal image patches combined with a local maximum operation for feature extraction can significantly improve recognition performance.[1]

It is a field of research in Pattern Recognition, Artificial Intelligence and Computer Vision. The aim of this research is to actualize the different methods of offline handwriting recognition like, Support Vector Machine, Artificial Neural Network, Hidden Markov Model etc. in the fields of HCR and OCR .[4]

Character recognition (CR) has been widely studied in the previous half century and develops to a level sufficient to produce technology determined applications. Nowadays, the precise recognition of machine printed characters is considered mainly a solved trouble. A lot of commercial products are paying attention towards that way, achieving high recognition rates. At the recognition phase a features are extracted from Characters in order to classify them to predefined classes techniques. However, handwritten character recognition is reasonably difficult. So, handwritten recognition documents is still a subject of active research Now, the quickly growing -computational power enables the execution of the present CR methodologies and creates a rising demand on many promising application field, which require more highly developed methodologies.[5]

We can compare the different classification performances obtained with sparse coding, Gabor wavelets, and principle component analysis. We conclude that the learning of a sparse representation of local image patches combined with a local maximum operation for feature extraction can significantly improve recognition performance.[3]

III.PROPOSED METHODOLOGY

As of late, o create subjective and social abilities for understudies , instructive mechanical autonomy has pulled in the high enthusiasm of educators and specialists as a significant apparatus . It has been broadly produced for understudies from various survey that to visually recognise data either in written form or printed a special system is designed to improve the working efficiency of the existing system and reduce the maintenance cost along with the working cost .

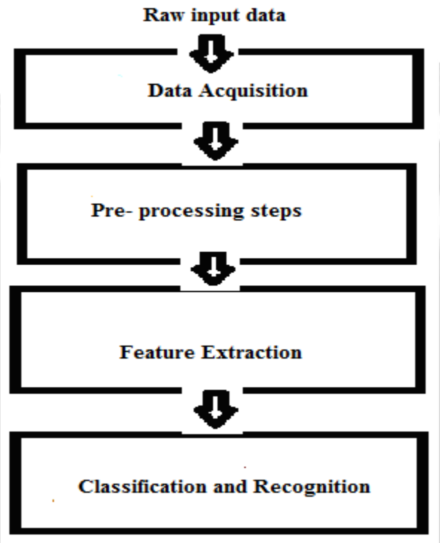


Fig1.2 Data Processing in OCR

Through the outline,creation, and a linearly supported vector machine is used to develop the system to recognise data and generate a feasible digital output from fig.1.2 . By utilizing it, the understudy will attempt to take in the present patterns in ocr innovation which is identified with the utilization of HCR, OCR, , SVM , k-nn model and Markov mode along with python based programming. It will be a perfect progression show thestudy the mechanism of an OCR Sytem, which incorporates existing feature and installed support vector machine . These abilities make it likewise has a potential in tackling numerous difficulties in industry and education centres to collect data digitallly. In this paper we propose a computer based programming that not only utilises the features of the provided data but also improvise it by removing errors . In this manner, for fruitful working of ocr and a lot of data can be saved digitally mention clearly fig1.3.

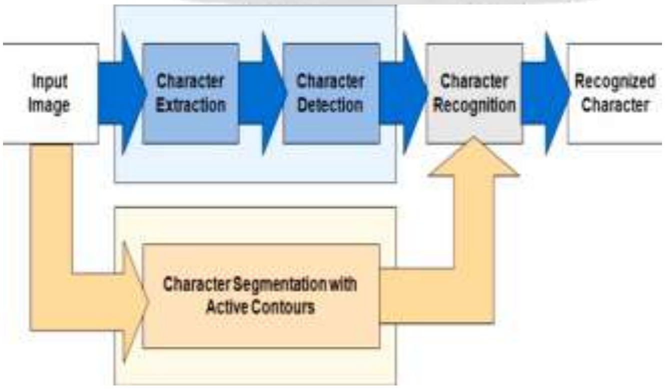


Fig1.3 Workflow in OCR

IV. DESIGN AND IMPLEMENTATION

The first thing we have to do is put Input image and then start the process . If there is already a processed image which directly go to the binarization and proceed with the slent correction and smoothing .Now we remove the noise for all the improving the quality of the image or else font . Now , after the aquiring of the data and then extracting features we normalize the size of fonts . With help of the very use of the Linear Vector Machine Classifier we then transformation of extremely complex data to figure out what and how to separate the data which is based on the labels or outputs is defined.In the feature extraction stage the each character is represented as a feature of vector, which then becomes its identity. The major goal of this feature extraction is to the extract a set of features, which then maximizes recognition rate with atleast amount of elements .Set up to the page and generate then required page for storing data . Save the data image and process is then completed .

1.Image Manipulation

Electronic questionnaires then can be used sent to the specialist operators then back to original operator if then necessary with the electronic questionnaires the same use of questionnaire can be done worked on the simultaneously by two or more persons . Electronic questionnaires are then readily available for the post census analysis (easier access to the questionnaires) .Parts of the various questionnaires on the screen at once for the inter record editing .

2. Binarization

Refers to conversion of the gray-scale image into the binary image. Two categories of the thresholding fig1.4:

1.Global, picks of one threshold value for the entire document of the image which is often based on an the estimation of a background level from intensity of a histogram of the image.

2.Adaptive (local), uses different values for each pixel according to a local area information.

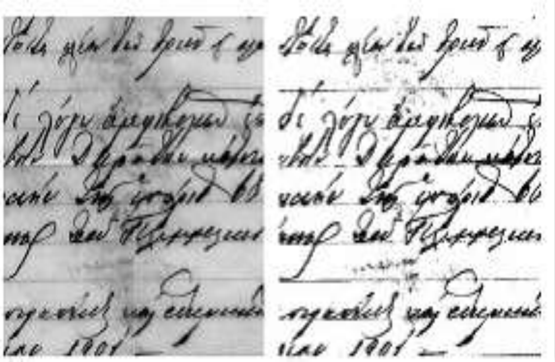
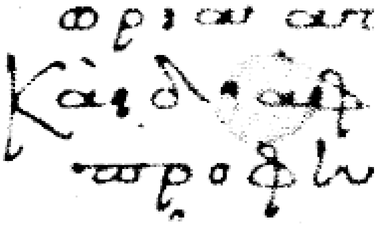


Fig1.4 Binarization

3. Noise Reduction

Noise reduction also improves the quality of the document.The two main approaches: (i)Filtering (masks) (ii)Morphological Operations (erosion, dilation, etc) in fig1.5.



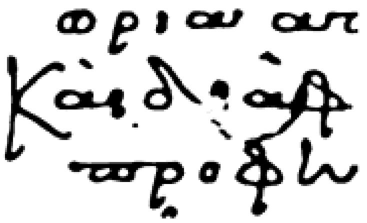


fig1.5 Noise reduction

4.Skew Correction & Slant Removal

Skew Correction methods are used to align the paper document with the coordinate system of the scanner whereas slant removal methods are used to normalize the all characters to a standard form.

Methods for slant removal are:

- Bozinovic – Shrihari Method (BSM)
- Calculation of the average angle of near-vertical elements

5. Linear SVM(Support Vector Machine):

Linear SVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set. Our comparisons with other known SVM models clearly show its superior performance when high accuracy is required. SVM-A Support Vector Machine is a supervised machine learning algorithm which can be used for both classification and regression problems.

V. MATERIAL AND METHOD

- System : Pentium IV 2.4 GHz
- Hard Disk :250 GB.
- Monitor :15 VGA Color.
- Ram :4 GB

1. Python 3.6.2

Python is a widely used high-level programming language for general-purpose programming . Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including object-oriented, imperative, functional programming, and procedural styles. It has a large and comprehensive standard of library.Python interpreters are available for many operating systems, allowing Python code to run on a wide variety of systems.

2. NumPy

It is library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays from fig1.6.

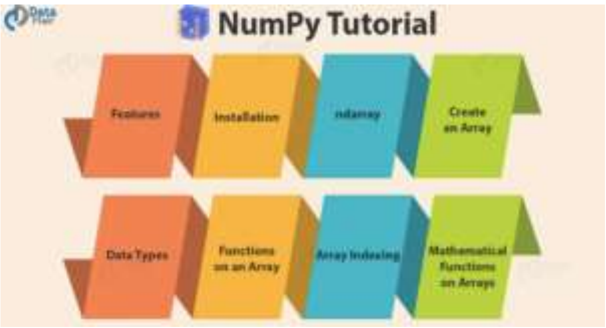


Fig1.6 Numpy

3. Anaconda Platform

Anaconda® is a package manager, an environment manager, a Python distribution, and a collection of over 1,000+ open source packages. Its free to install and easy to use , and it offers free community support.Over 150 packages are automatically installed with Anaconda.Over 250 additional open source packages can be individually installed from the Anaconda repository with the condainstall command from fig1.7.

Fig1.7 Anaconda Platform

VI. RESULT AND DISCUSSION

This method of digit recognition needs a maximum degree of correctness when proposed system working on a very huge amount of activity of counting lots of handwritten text which may not be possible manually as humans tend to get exhausted due to dull nature of job and will need a great continuous concentration from a man for a long period of time . A new technique of Linear Support Vector Machine for Digit Recognition. Linear SVM generates less recognition times as well improves recognition made in this dissertation to develop an accurate and digit recognition system. This work demonstrates better performance as compare to correlation method which is one of efficient techniques for matching.

VII. CONCLUSION & FUTURE WORK

In this project we proposed a novel algorithm for handwritten digit recognition. The goal was to use simple feature set as input for support vector machine that was used for classification. Further developed for alphabets, which will make this Digit Recognition more feasible. A better Data-Set will increase the accuracy and prediction. Optimal SVM models were determined by recent swarm intelligence algorithm, bat algorithm. Bat algorithm was adjusted and used for parameter tuning of the support vectormachine.We tested our proposed method on standard MNIST dataset and achieved global accuracy of 95.60%.

REFERENCES

1.Machine Learning Final Project: Handwritten Sanskrit Recognition using a Multi-class SVM with K-NN Guidance .
[1] http://sanskritlibrary.org/. 1

[2] http://www.indsenz.com/int/. 3

[3] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. Advances in neural information processing systems, pages 831– 837, 2001. 6

2. Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study.

[1] ARCHANA A. SHINDE, D**.** 2012.Text Pre-processing and Text Segmentation for OCR. International Journal of Computer Science Engineering and Technology, pp. 810-812.

[2]ANAGNOSTOPOULOS,C.,ANAGNOSTOPOULOS, I., LOUMOS, V, & KAYAFAS, E. 2006. A License Plate Recognition Algorithm for Intelligent Transportation System Applications.**.**, IEEE Transactions on Intelligent Transportation Systems, pp. 377- 399.

[3] Y. WEN, Y. L. 2011. An Algorithm for License Plate Recognition Applied to Intelligent Transportation System., IEEE Transactions on Intelligent Systems, pp. 1-16.

3. Simple Method for High-Performance Digit Recognition Based on Sparse Coding .

[1] D. Keysers, C. Gollan, T. Deselaers, and H. Ney, “Deformation models for image recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 8, pp. 1422–1435, 2007.

[2] K. Fukunaga, Introduction to statistical pattern recognition (2nd ed.). San Diego, CA, USA: Academic Press Professional, Inc., 1990.

[3] J. G. Daugman, “Complete discrete 2-D gabor transforms by neural networks for image analysis and compression,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36, no. 7, pp. 1169–1179, 1988.

4. Review of Offline Handwriting Recognition Techniques in the fields of HCR and OCR

[1] Beigi, H.S.M., “An overview of handwriting recognition”, Proc. The 1st Annual Conference on Technological Advancements in Developing countries, 1993. Columbia University. p. 30-46.

[2] Er. Neetu Bhatia, ”Optical Character Recognition Techniques: A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014

[3] Ranyang Li, Hang Wang, KaifanJi, “Feature extraction and identification of handwritten characters”, 8th International Conference on Intelligent Networks and Intelligent Systems, 2015

5. Different Classificaation Techniques for OCR

[1] M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis and Machine Vision (PWS Publishing, Books/Cole Pub. Company, 2nd Ed, 1999).

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Ha®ner, Gradient-based learning applied to document recognition, Proc. IEEE, 86(11): 2278-2324, 1998.

[3] C.Y. Suen, K. Kiu, N.W. Strathy, Sorting and rec-ognizing cheques and ¯nancial documents, DocumentAnalysis Systems: Theory and Practice, S.-W. Lee andY. Nakano (eds.), LNCS 1655, Springer, 1999, pp. 173-187.