International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 5, Issue 1, 2018, PP 9-16
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0501002
www.arcjournals.org

Optical Recognition of Digital Characters Using Machine Learning

Dr Sunanda Dixit[1], Bharath M[2], Amith Y[2], Goutham M L[2], Ayappa K[2], Harshitha D[2]

[1]*Associate Professor, Department ISE, Dayananda Sagar College of Engineering, Bengaluru, India*

[2]*Student, Department ISE, Dayananda Sagar College of Engineering, Bengaluru, India*

***Corresponding Author: Dr Sunanda Dixit**, Associate Professor, Department ISE, Dayananda Sagar College of Engineering, Bengaluru, India*

Abstract: *Optical Character Recognition (OCR) plays an important role in document image processing. Recognition of characters in a smart way is gaining importance in the modern days, as huge piles of data is generated, and it needs to be processed and manipulated. OCR eliminates the need of retyping in case of erasure of an important digital file accidentally, it quickens digital searches, enables faster text editing, saves up the storage space with the use of efficient algorithm. Accessibility of OCR to the visually impaired is a great progress, Where users can scan books, magazines, incoming mails or other programs. OCR can also be coupled with a voice-over utility, or voice synthesizer, that reads out the text recognized optically by the software. OCR also has its own bottlenecks. It may not recognize complex text, mathematical variables, hand written scripts. High end OCR utilities are used in areas where required, but it comes with an additional computational cost. This paper aims at presenting an OCR utility which recognizes text characters, using a machine learning model.*

Keywords: *Artificial intelligence, classification algorithm, machine learning, Optical character recognition, machine learning*

1. INTRODUCTION

Identification of optically processed characters is know as character recognition (OCR). OCR is a technique used to process variety of documents, PDF or digital images into American Standard Code for Information Interchange (ASCII) or other equivalent machine editable form which the data can be edited or searched.

Recent improvement in pattern recognition by many applications has been demanding , such as OCR, classification of Document, Data Mining etc. Use of OCR has vital role in Document scanners, character recognition, language recognition, security, authentication in Bank etc. OCR is classified into two types: online character recognition and offline character recognition system. Online OCR out beats offline OCR as characters are processed as it is written, this avoids initial stage of identifying the character .Offline OCR are further sub-divided into printed and handwritten OCR. In offline OCR are processed typically by scanning the typewritten /handwritten characters into binary or gray scale image to the recognition algorithm.

With advancement of OCR scanned documents are more valuable just then normal image file, turning into textcontents which are recognized by computers.OCR finds a better way of autom atically entering into electronic database over conventional method of manually retyping. Usual problem with OCR lies with segmentation of characters or symbols which are joined. Input image is directly proportion to the accuracy of the OCR. Yet no recognition algorithm competes with the quality of the human intelligence, it is still proven to much faster which is still attractive.

2. LITERATURE SURVEY

Manuscripts come in different shapes and sizes. They vary in the type of fonts, style of the text, the punctuations etc. Words in a text pattern may often be connected, which are segmented to pose them as logically separate entity. Application of such text recognition is of utmost importance in areas like post offices, schools etc. Where the man power is diminishing now a

days [1]. Besides, handwritten text in various languages or scripts can also be optically recognized optically. One such example is, Hindi language can be recognized by the structural description (features) of the characters. Features roughly translate to density, segment and moment features of the characters [1]. Other Indian scripts like Kannada and Devanagri are recognized by a contour based network, which focuses its attention on the contours/boundaries of the characters. Later, an algorithm is formulated based on the feature extraction, to which other characters to be optically recognized are input.[1]

Normalization of the input is a pivot step for the prediction model, and Aspect ratio adaptive normalization (ARAN) is one such efficient planes each plane is divided into identical zones and intensity of each zone is recorded.[2]. A methodology for off-line handwritten character recognition is proposed in [4]. Feature extraction methods for handwritten characters and digits have been based mainly on two types of features (a) statistical derived from statistical distribution of points. (b) Structural the most common statistical features used for character representation are: (a) zoning, where the character is divided into several zones and features are extracted from the densities in each zone of the contour of the character by computing histograms of chain codes in each zone, (b) projections and (c) crossings. Character images are divided into uniform regions that are searched for vertical, horizontal and diagonal segments. The total number of such segments is fed to the classifier. Before employing the proposed feature extraction technique all character images must be black and white (b/w) and normalized to an N x N matrix.

Sunanda et al[3] proposed a method for text line segmentation which will be the input for kannada OCR recognition by using energy minimization problem which uses a new cost function.

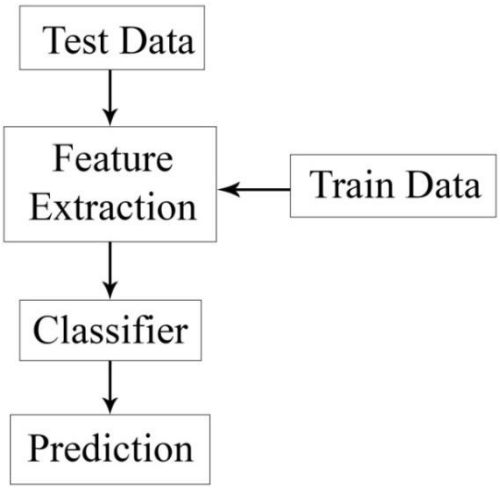
OCR work done on Indian culture scripts is proposed in [5]. In most of the Indian languages, a text line may be partitioned into three zones. The upper-zone denotes the portion above the head-line, the middle-zone covers the portion of basic (and compound) characters below head-line and the lower-zone is the portion below base-line. Traditionally, pattern recognition process are divided into template and feature-based approach. Early OCR systems employed only with template-based approach, but modern systems combine this with feature-based approaches to obtain better results. The feature-based approaches can be of two types, namely spatial domain and transform domain approaches. OCR can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications.

The aspect ratio mapping decides the efficiency of the ARAN model. in feature extraction segment, distribution of local strokes direction is most widely used method.[2] in ARAN, the image to be normalized is fit to a plane, with any of its one dimension filled depending upon its aspect ratio. Here, the plane dimension is assumed to be a square of length L. the aspect ratio is mapped by determining the width and height of the normalized image. The aspect ratio of a normalized image is scalable to that of original image [2]. Directional feature extraction can be of chain code feature, NCFE feature, and gradient feature. In NCFE, the contours and edges of the image are mapped to orientation. Recognition of English text using Energy minimization technique is proposed [7].

3. METHODOLOGY

The recognition system is input an image of any format(jpeg, png, etc). This is done either through scanning an image from any of the digital scanners or by loading it from the internal storage. Image processing involves several major functions performed on an image, out of which image pre-processing is one of the most vital and initial step[2]. Image pre-processing: there are various environments available to construct a machine learning model. Python is one such robust environment and r studio is also a powerful environment which is used to predict, plot and depict the information efficiently. We've used visual studio here which offers a variety of services. Image pre-processing involves getting the dataset i.e, the image which we work upon. Then, we import the required libraries. These libraries have some powerful built-in functions that help us to manipulate the data or visualize it in an efficient manner. Python offers some powerful libraries such as numpy - for high end arithmetic operations, pandas etc. Here, we import AFORGE library, whose framework includes the support for computer vision, image processing, video processing and Artificial Neural Networks,

fuzzy logic and other support. We then split the dataset into test set and training set. Train data is supposed to train the machine. It is on this data by which the machine learns. The test data is the data that is input to the machine to get results. After classification of data set into test and train sets, feature scaling needs to be done. It operates on the essential features of the image i.e. min bound box, segmentation[1] etc.



Block diagram of the model

3.1. Input

The input type to recognition system is of type scanned images .Which is of specific format such as JPEG, BMT etc . Input to the system is achieved through scanner, digital cameras or any other suitable digital input device. Preprocessing is usually done on input scanned image. It is important as it is required for segmentation. The input scanned images are converted into gray scale image before segmentation and bounding is performed on input image before segmentation. Bounding box is also known as Minimum bounding box, box with the smallest measure within which all the points lie.

Algorithm Steps:

1. Converting original image into grey scale image.

2. Bounding on input image

3. Draw the bounded box of input image

4. Fins minimum width of input image

5. To find Maximum height of input image

3.2. Segmentation

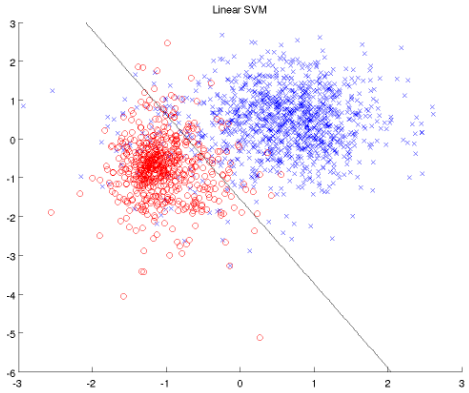
The recognition of Characters in an image optically usually has an image that is comprised by a sequences of characters. For our machine learning model to be efficient, we employ segmentation technique, which decomposes the sequence of characters in an image, to a sub image of individual character. Each sub image is resized uniformly to resolution of size 90*60 pixels for further processing.

3.3. Zoning/Feature Extration Method

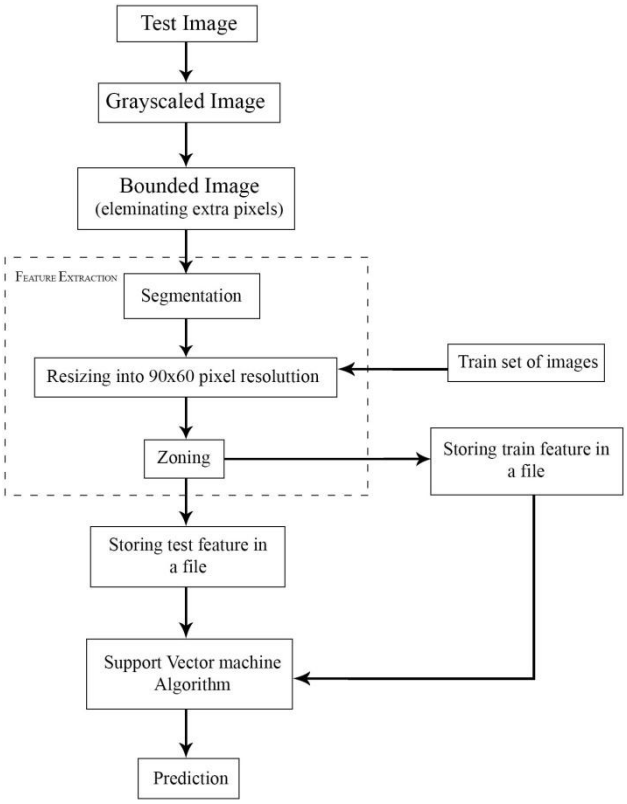
Feature Extraction/Feature Scaling/Zoning: in this section, the segmented images are subject to feature scaling. i.e., each image is examined thoroughly for its features and the data is recorded by the machine which later helps in a precise or accurate prediction[4]. There are various variants of feature extraction. Here, we employ diagonal feature extraction method, which is one of the efficient feature extraction methods for OCR. This method examines the diagonals of the input image and extracts the features present over there. It identifies the static and manuscripted characters. in segmentation, the images are resized to a resolution of 90*60 pixels, each of which is further classified to a resolution of 10*10 pixels each, amounting to a total of 54 uniform sized 'Zones' (hence the name Zoning). The features of each zone are extracted. In OCR, feature translates to the intensity of the black pixel in each zone the feature extraction method records the feature of each zone.

4. SUPPORT VECTOR MACHINE

Regression techniques are just a basic step towards creation of Machine Learning Model. They are capable of slicing and dicing the data efficiently, but fail to operate on a complex dataset, hence, SVM- support vector machine is used, which works on smaller data set, and is robust in building a powerful Machine Learning Model. SVM is used for both classification and regression of dataset. It use is predominant in classification. A plotting function is used here, which visualizes the value of each data feature as a coordinate in a coordinate plane. Then, classification is done by a hyper plane. Python offers Scikit-learn library to implement SVM. The SVM class is import and then a regress or is created, which is late used to fit the model and visualize the data results. Here, we use the method Train Classifier() to read the train samples and pass the parameters to it. Gamma parameter is the kernel coefficient for rbf, poly, sigmoid values, the default being rbf. The value of gamma is directly proportional to accuracy of our Machine Learning Model. The function Testclassifier () predicts the test results.



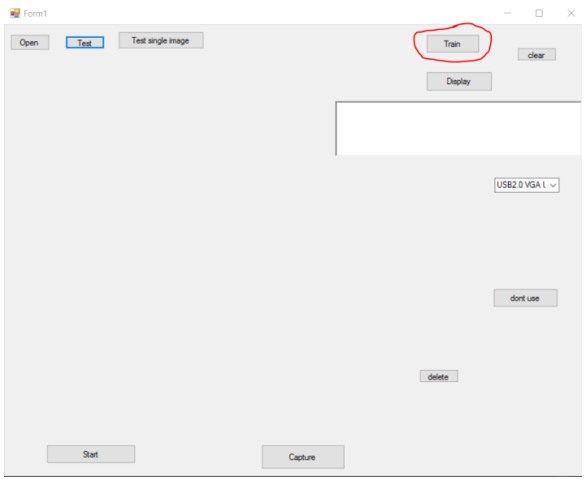
Visualiztion of SVM classifier[5]



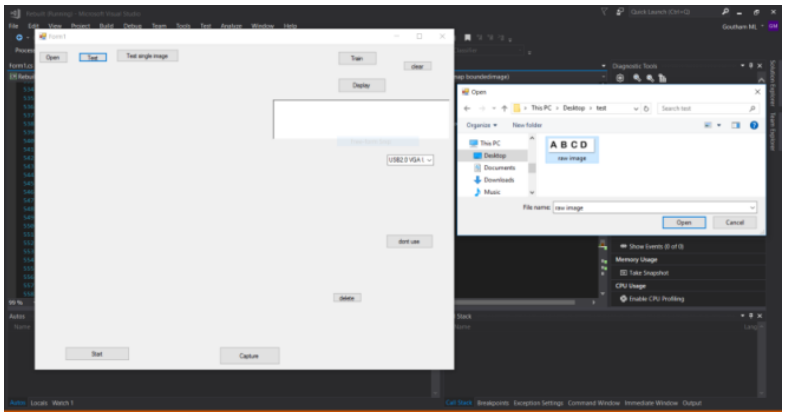
Flowchart of the model

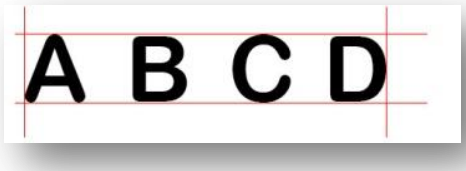
5. OUTPUT

1. Training the classifier with data sets containing more than 1000 images of each characters.



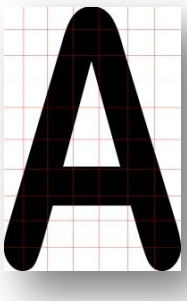
2. Input image is converted into gray scale and cropped to get bounded image

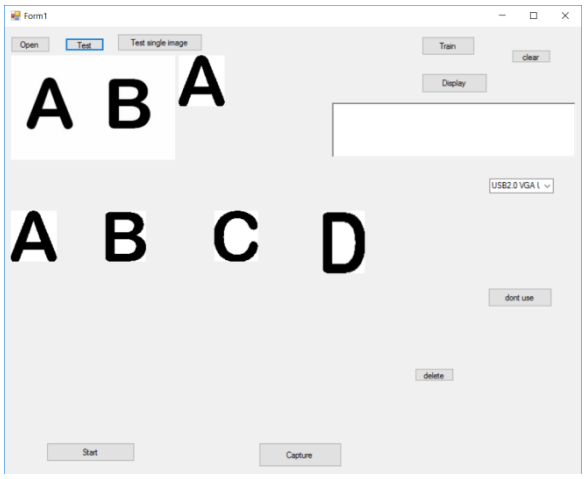




Cropped image is segmented into sub images of single character.

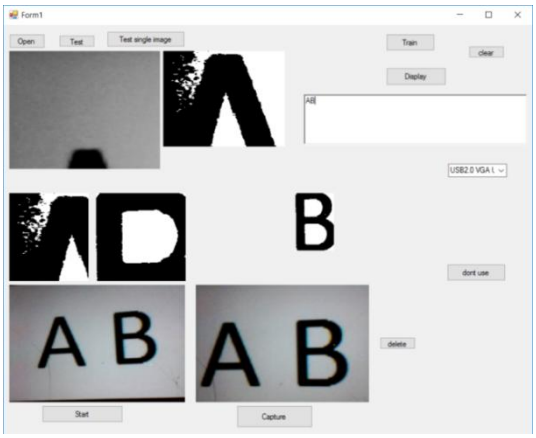
3. Each sub image is resized to resolution of 90x60 pixels and passed to feature extraction. At this phase whole sub image is divided into 54 zones where each zone containing 10x10(100 pixels). Let us consider sub image of ‘A’.





4. SVM classifier performs classification of test data against train data. Predicted results are read from the file 'result'. Texts are displayed inside text box and it is read out.

5. Character recognition can be performed on printed texts is done by capturing image and following above



6. RESULT

Detailed implementation of OCR is presented above. This model is able to recognize texts in optical form. Input can be fed to this model either through scanning printed text or a digital image. Texts will be printed in text box & each character is read out by the voice synthesizer function. This will be helpful for visual impaired people. This model can be extended to recognize alphabets of various languages. Also the model can be trained to recognize handwritten characters which will reduce man power in various organization. Prediction accuracy can be increased by training with more number of train images.

REFERENCES

[1] Cheng-LinLiu, Kazuki Nakashima, Hiroshi Sako, Hiromichi Fujisawa, "Handwritten digit recognition: investigation of normalization and feature extraction techniques", Pattern Recognition, Elsevier, Volume 37, Issue 2, PP. 265-279, February 2004

[2] Hanmandlu, O.V. Ramana Murthy, Fuzzy model based recognition of handwritten numerals, Pattern Recognition, Elsevier, Volume 40, Issue 6, PP.1840-1854, June 2007

[3] Sunanda Dixit, Suresh Hosahalli Narayan, Mahesh Belur, Kannada text line extraction based on energy minimization and skew correction, Advance Computing Conference (IACC), IEEE International, Pages62-67,2014

[4] Georgios Vamvakas, Basilis Gatos, Stavros J., Perantonis, Handwritten character recognition through two-stage foreground sub-sampling, Pattern Recognition, Elsevier Volume 43, Issue 8, Pages 2807-2816, August 2010

[5] U. Pal, B.B. Chaudhuri., Indian script character recognition: a survey, Pattern Recognition, Elsevier, Volume 37, Issue 9, Pages 1887-1899, September 2004.

[6] https://cw.fel.cvut.cz/wiki/_media/courses/ae4b33rpz/labs/08_svmocr_svm_trn.png?w=600&tok=834ae2

[7] Kanchan Keisham and Sunanda Dixit, Recognition of Handwritten English Text Using Energy minimization, Springer Information Systems Design and Intelligent Applications , Part of the Advances in Intelligent Systems and Computing book series , AISC, volume 435), pp 607-614, 2016.

AUTHOR'S BIOGRAPHY

Professor Sunanda Dixit, earned her Doctorate in Computer and Information Science Engineering with a special focus on Digital Image Processing from the renowned Visvesvaraya Technological University, a state university in Belagavi, Karanataka in 2015. She started her career in industries from 1996. She moved to academic career field from 2009 and served in various Engineering Colleges. She has published more than 60 research papers to her credit in referred and indexed journals, book chapters and conferences at international and national levels. It includes a book titled "Optimize Bandwidth for Signaling Protocols by Compression Technique" which has been published in Lambert Academic Publishing at Germany 2012.

In addition to that, she is the editorial board member of various International journals, Providing her services in various Technical Programme Committee member, Reviewer for various international and national conferences and National Advisory member for international conferences.

Bharath M, under graduate student in Information Science and Engineering. He has been working on Artificial Intelligence thorough machine learning. He want to explore more on it.

Amith Y, under graduate student in Information Science and Engineering. He is an emerging data science enthusiast, with interests in Machine learning and deep learning and statistical analysis, using Excel and SAAS.

Goutham M L, under graduate student in Information Science and Engineering. He is passionate about software development and web development. He was into web development (backend) until he was exposed to AI through Machine learning. He is now learning Deep learning and want to explore and learn more in it. He also work as freelance graphic designer.

Ayappa K, under graduate student in Information Science and Engineering. He has worked on web and android app development. He has completed an internship on web development in a startup. He is am familiar with c,c++ and java programming languages.

Harshitha D, under graduate student in Information Science and Engineering. She is interested to work on AI and Machine learning. Interested in web development and currently interning as web developer in a stratup.

Citation: *Dr Sunanda Dixit et.al (2018). Optical Recognition of Digital Characters Using Machine Learning, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 5(1), pp.9-16, DOI: http://dx.doi.org/10.20431/2349-4859.0501002*