Recognition of Hand Written And Printed Text Using ANN

Shiny Ronisha. T[1], Ms. Thayanandeswari.C.S[2], Seline Manju.R[3]
[2, 3]Assistant Professor
[1, 2, 3] PET Engineering College

Abstract- *Character recognition still remains an active area for research towards exploring the new techniques that would help in improving recognition accuracy. This project focuses on recognition of character from the image with the help of ANN. The steps involved in this methodology are Preprocessing, Segmentation, Feature Extraction using CPRFS, Training, Classification and Recognition using ANN. Preprocessing is a series of operation performed on input image. Next the preprocessed image is segmented into isolated character by assigning a number to each character using labeling which provides information about number of character in the image. Feature Extraction using CPRFS it extract different line type that form a particular character and the gradient measures the magnitude and direction of the greatest change in each pixel. The Data sets, containing texts are used to train the system. Finally the ANN classifier is used to classify the words and the features of each character written in the input are extracted and then passed to the neural network and text characters are recognized. The ANN proposed recognition system gives high level accuracy*

Keywords- Handwritten, printed text pre-processing, segmentation, feature extraction, recognition, optical character recognition, Artificial Neural Network

I. INTRODUCTION

Handwritten character recognition is a field of research in artificial intelligence, computer vision, and pattern recognition. A computer performing handwriting recognition is said to be able to acquire and detect characters in paper documents, pictures, touch-screen devices and other sources and convert them into machine-encoded form. Its application is found in optical character recognition and more advanced intelligent character recognition systems. Most of these systems nowadays implement machine learning mechanisms such as neural networks. Machine learning is a branch of artificial intelligence inspired by psychology and biology that deals with learning from a set of data and can be applied to solve wide spectrum of problems.

A supervised machine learning model is given instances of data specific to a problem domain and an answer that solves the problem for each instance. When learning is complete, the model is able not only to provide answers to the data it has learned on, but also to yet unseen data with high precision. Neural networks are learning models used in machine learning. Their aim is to simulate the learning process that occurs in an animal or human neural system. Being one of the most powerful learning models, they are useful in automation of tasks where the decision of a human being takes too long, or is imprecise.

A neural network can be very fast at delivering results and may detect connections between seen instances of data that human cannot see. We have decided to implement a neural network in an Android application that recognizes characters written on the device's touch screen by hand and extracted from camera and images provided by the device. Having acquired the knowledge that is explained in this text, the neural network has been implemented on a low level without using libraries that already facilitate the process. By doing this, we evaluate the performance of neural networks in the given problem and provide source code for the network that can be used to solve many different classification problems. The resulting system is a subset of a complex OCR or ICR system; these are seen as possible future extensions of this work. It is a field of research in pattern recognition, artificial intelligence and machine vision. Though academic research in the field continues, the focus on character recognition has shifted to implementation of proven techniques.

The organization of the paper is as per the following. In section II literature survey is explained. Section III explains the proposed methodology. Section IV illustrates the dataset. Section V shows the conclusion and future enhancement.

II. LITERATURE REVIEW

Handwritten Character Recognition [HCR][1] is a general procedure of written texts or digitizing pictures of printed with the goal that they could be electronically revised, put away and looked through more proficiency and accurately. The target of an OCR technique is an acknowledgment of

composition (same as people) in a troublesome article. OCR techniques are significantly classified in two sorts, First online text recognition and Second offline recognition. Offline OCR is considered in two subclasses initially is typed and second is handwritten text.

Optical character recognition is a technique for consequently recognizing various character from a record picture moreover gives complete alphanumeric recognition for printed or handwritten characters, text numerical, letters, and symbols into PC procedure capable format including ASCII, Unicode thus forth [2]. These OCR innovations help to look at interesting records written in English, Chinese, Hindu, Arabic, Russian and other languages. This work illustrates the audit of certain investigates has been made in English, Arabic, and Devanagari characters The initial step is pre-processing, for example, noise detection and removal, binarization, etc. The segmentation of archives images into line, word, and characters. This is trailed by feature extraction for representing character pictures and a classification module, finally post-processing.

Segmentation of handwritten archive pictures into text lines and words is a basic task for Optical character recognition[3]. However, a feature of written text is unpredictable and differs from person to person. To address the issue, the authors have characterized the word segmentation issue as a binary quadratic assignment task that considers the pairwise relationship between the spaces similarly as the probabilities of individual space.

Printed text recognition is the limit of PCs to understand input characters from an outside source for example e-forms and e-reports[4]. Machine learning algorithms gives an ability to decipher the printed text to digitized characters. The purpose is to actualize a machine written character recognition framework for lower case English characters for two unique styles i.e.., Times New Roman and Arial by using the Learning Vector Quantization algorithm.

Segmentation of handwritten text archives into singular character or digit is a critical stage in archive examination, character recognition, and numerous different zones. In this paper, the author has examined different strategies to segment a picture containing text at different degrees of segmentation[5]. Text segmentation and histogram technique to partition each character individually.

Handwritten recognition is the limit of PC's to perceive and decode written text from sources[6]. Recognition of handwritten characters by a PC is a troublesome issue because of the human handwriting changeability, uneven skew, orientation writing habit, style. In this work, the test characters are sorted into vowel/consonant classes utilizing the multiclass SVM classifier, acquiring 89.84% accuracy in recognizing Kannada vowels and 85.14% accuracy in recognizing consonants.

Optical Character Recognition [OCR] of printed Latin content documents are universally guaranteed as a tackled issue[7]. The latest methodologies recognize characters by segmentation. The paper portrays a text line recognition approach utilizing multi-layer perception [MLP] and hidden Markov models [HMMs].

This paper expected to prepare our classifier in case we are considering using data mining methods for such purposes[8]. There are a few set up generic classification techniques that can be utilized together with feature extraction mechanism yet it is imperative to know which of them improve under which condition. This assesses three methodologies for OCR from manuscripts and considers their outcome.

The paper proposed a novel multi-model archive image recovery system by misusing the data of text and design areas[9]. The system applies various part based hashing formulation for the generation of composite records utilizing various modalities. In the ensuring commitment propose novel multi-modular archives report ordering system for recovery of an old and degraded text document by joining OCR'ed content and image-based representation utilizing learning.

The paper proposed the utilization of stacking denoising encoder for programmed feature extraction clearly from raw pixel estimation of pictures[10]. Such profound learning systems have not been applied for recognizing the Urdu text so far. Along these lines, prepared systems are approved and tried on degraded forms of UPTI informational collection.

Character recognition assumes a significant role in removing the necessary content from the document[11]. The essential point is to perceive the printed characters in a given data picture and isolating it. The machine learning procedure is utilized, where the framework is first arranged for all the letters in order and evaluates the English language along with the expected result. There are four stages in this particular technique pre-processing, segmentation (Line segmentation, Word segmentation, character segmentation) and next is to identify the features of each character. And lastly, the classification has been done.

In the current strategy, it makes use of topographical features and projection profiles to identify character segmentation region from monochrome images [12]. Using a multi-stage graph search algorithm, a nonlinear character segmentation routes is found. Lastly, Recognition based segmentation technique is utilized to check the correctness of the nonlinear character segmentation routes and output. This approach is found to be successful in identifying overlayed and close by characters.

Modi was helpful content in the kingdom of medieval Maharashtra[13]. In the region of incomparable Maharatha Chatrapathi Shivaji and the rule of Peshwas, the content was broadly joined in administering the state. This content was very like the shorthand, Around then it was utilized in Maharashtra to set up the document, for example, Property issues, Dan-Patra, Land revenue, and so on, right now Modi content is considered, Recognizable proof and recognition of handwritten of Modi characters are done. The database of handwritten tests set up by utilizing the ANSEP program

In this paper has designed a very distinctive android based Multilanguage smart device application that improves user's writing experiences[14]. The main purpose of the work is to perceive character recognition algorithms that perform better on low-performance devices. Right now strategies are language free and have been effectively utilized for many languages of India.

This paper inspects the issues in recognizing the Devanagari characters in the wild like sign sheets, commercial, logos, and soon[15]. It manages the issues in recognizing the machine printed and the handwritten Devanagiri characters. The current OCR method is useful for the scanned images of printed text, perform ineffectively on characters extracted from the images of the wild. It is a result of the images of the wild contain the unforeseen fonts, 3D impacts, and have distortion and noise in characters. character recognition, and many recent methods have been proposed to design better feature representations and models for both. In this paper, we apply methods recently developed in machine learning–specifically, large-scale algorithms for learning the features automatically from unlabeled data–and show that they allow us to construct highly effective classifiers for both detection and recognition to be used in a high accuracy end-to-end system.

III. PROPOSED SYSTEM

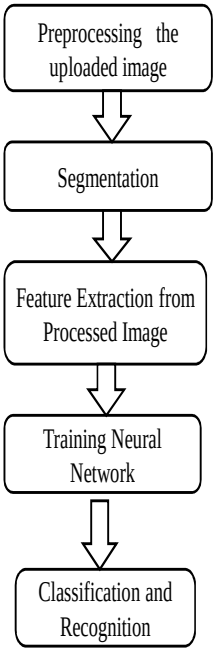
The scope of this design document is to achieve the features of the system such as pre-process the images, feature extraction, segmentation and display the text present in the image.

An Artificial Neural Network (ANN) is an information-processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems.The abilities of different networks can be related to their structure, dynamics and learning methods. Neural Networks offer improved performance over conventional technologies in areas which includes: Machine Vision, Robust Pattern Detection, Signal Filtering, Virtual Reality, Data Segmentation, Data Compression, Data Mining, Text Mining, Artificial Life, Adaptive Control, Optimization and Scheduling, Complex Mapping and more.

The proposed methodology uses some techniques to remove the background noise, and features extraction to detect and classify the handwritten text. The proposed method comprises of 4 phases:

1. Pre-processing.
2. Segmentation.
3. Feature Extraction.
4. Classification and Recognition.

Block diagram of proposed method



The architecture diagram shows the different stages of text recognition of handwritten and printed text.

Handwritten or printed papers are scanned to obtain input to the recognition system. The scanned input is saved for further processing in an image format (.jpg) as shown in Fig2.

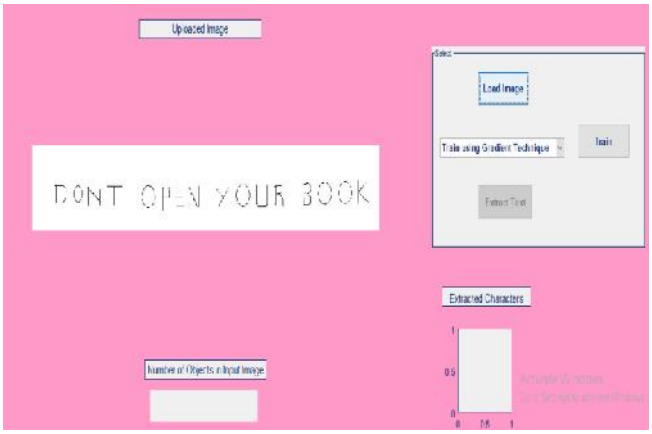


Fig:3.2 Input image

Pre-processing is the first step in image processing and it is one of the most important steps in pattern recognition. Processing a colored image increases the complexity and also reduces the speed of processing. So, the input image is converted to a grayscale image. Pre-processing enhances the details of the image by removing distortion/ noise and preserving the edges. This is achieved by applying a Median filter. The segmentation process starts after preprocessing.

Binarisation of an image converts it into an image which only have pure black and pure white pixel values in it. Basically during binarization of a grey-scale image, pixels with intensity lower than half of the full intensity value gets a zero value converting them into black ones. And the remaining pixels get a full intensity value converting it into white pixels.

Segmentation modifies the representation of an image into a much simple and relevant form to analyze further. Image segmentation is usually used to locate and discover the exact position of characters and boundaries like lines, curves, etc. It is an operation that seeks to decompose an image of sequence of characters into sub images of individual symbols. Character segmentation is a key requirement that determines the utility of conventional systems. Different methods used can be classified based on the type of text and strategy being followed like straight segmentation method, recognition-based segmentation and cut classification method.

The gradient measures the magnitude and direction of the greatest change in intensity in a small neighbourhood of each pixel. (In what follows, "gradient" refers to both the gradient magnitude and direction). Gradients are computed by means of the Sobel operator. The Sobel templates used to compute the horizontal (X) & vertical (Y) components of the gradient

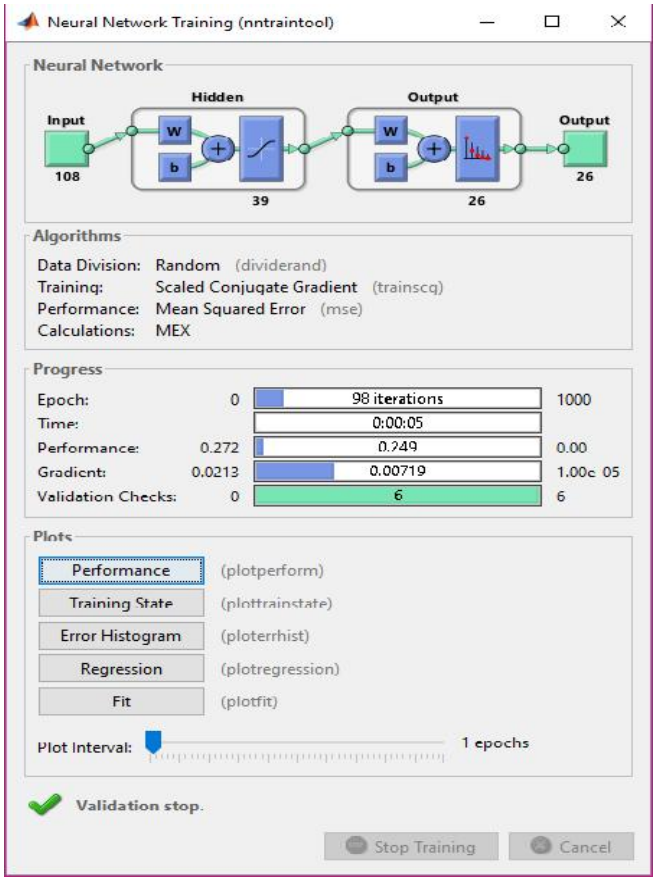


Fig:3.3ArtificialNeuralNetwork

Fig 3.3 shows theNeural network in a training process the training set is a set of pairs of input patterns with corresponding desired output patterns. The original image is compare to the training image to produce the output image. It have three different layers are input, output and hidden layers. The hidden layer perform nonlinear transformations of the inputs entered in the network.

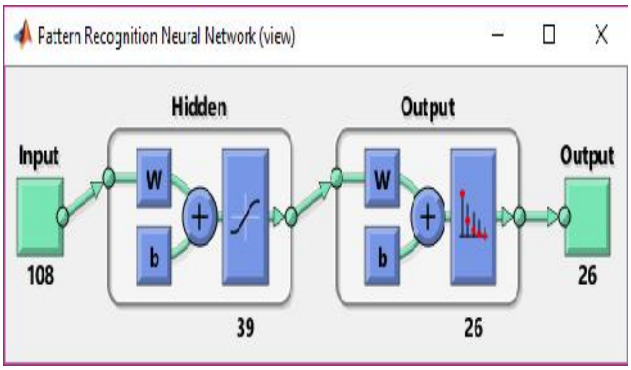


Fig:3.4Patternrecognition

The process that takes in raw data and makes an action based on the category of the pattern.It find the regularities and similarities in data.During training, the network is trained to associate output with input patterns and when the network is used, it identify the input pattern and tries to output the associated output pattern.

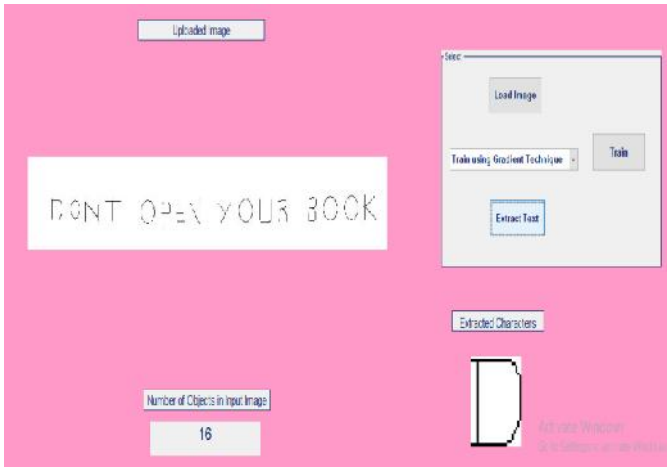


Fig:3.5Character recognition

After recognition the input image containing noise is segmented and the noise can be removed using morphological operations.In dilation operation it adds pixels to the boundaries of objects in a image and the area of foreground pixels grow in size and holes within those regions become smaller. Next the erosion operation performed to removes pixels on object boundaries and the area of foreground pixels shrink in size and holes within those areas become larger.

Fig.3.5 shows the character extracted from the input image compared with the trained data set.once the data set is trained and the character extracted randomly.

IV. DATA SET

The character data set consists of four different Styles of a printed and handwritten text document. The styles are Comic- Sans, Calibri, Tahoma, and San-Serif. Handwritten or Printed text is scanned to obtain the input for the recognition system. For the training image, Read the image which is in jpg format, and the image should be resized to 256*256 pixels. Once pre- processing, bounding box, splitting characters , and feature extraction are completed; the name of the style will be recognized.

V. CONCLUSION AND FUTURE ENHANCEMENT

Thus the postal automation has been recently integrated into the research agenda of the pattern recognition and computer vision communities. Here we exploit proximity cues in order to describe the investigated regions on envelopes. We propose two proximity descriptors encoding

spatial distributions of the connected components obtained from the binary envelope images. To locate the destination address block, these descriptors are used together with cooperative profit random forests (CPRFs). Finally the ANN classifier is used to classify the words. The features of each character written in the input are extracted and then passed to the neural network. Data sets, containing texts written by different people are used to train the system. The proposed recognition system gives high levels of accuracy as compared to the conventional approaches in this field.

The proposed recognition system gives high levels of accuracy as compared to the conventional approaches in this field. Neural network followed by the Back Propagation Algorithm which compromises Training. An experimental result shows that ANN with back propagation network yields good recognition accuracy of 99%.In future work A Convolutional Neural Network (CNN) is trained on the resulting dataset to predict the probability of each pixel of being foreground given a patch centered on it. The CNN learns what a finger vein pattern is by learning the difference between vein patterns and background ones. The pixels in any region of a test image can then be classified effectively.

REFERENCES

[1] Sonkusare, M., & Sahu, N. (2016). A survey on handwritten character recognition (HCR) techniques for English alphabets. *Adv Vis Comput Int J*, *3*(1), 1-12.

[2] Awel, M. A., & Abidi, A. I. (2019). Review on optical character recognition. *no. June*, 3666-3669

[3] J. Ryu, H. I. Koo and N. I. Cho, "Word Segmentation Method for Handwritten Documents based on Structured Learning," in IEEE Signal Processing Letters, vol. 22, no. 8, pp.1161 -1165,Aug.2015.doi: 10.1109/LSP.2015.2389852.

[4] Shirzadi, E. (2016). Machine-Written Character Recognition Using A Supervised Machine Learning Approach. *J. Elec. Commu. Eng. Resol*, *1*(1), 6-10.

[5] Dave, N. (2015). Segmentation methods for hand written character recognition. International journal of signal processing, image processing and pattern recognition, 8(4), 155-164.

[6] Angadi, S. A., & Angadi, S. H. (2015). Structural Features for Recognition of Hand Written Kannada Character based on SVM. International Journal of Computer Science, Engineering and Information Technology, 5(2), 25-32.

[7] S. F. Rashid, F. Shafait and T. M. Breuel, "Scanning Neural Network for Text Line Recognition," 2012 10th IAPR International Workshop on Document Analysis Systems, Gold Cost, QLD, 2012, pp.105 - 109.doi: 10.1109/DAS.2012.77.

[8] S. H. Tanvir, T. A. Khan and A. B. Yamin, "Evaluation of optical character recognition algorithms and feature extraction techniques," 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, Ireland, 2016, pp. 326 -331.

[9] E. Hassan, S. Chaudhury and M. Gopal, "Multi-modal Information Integration for Document Retrieval," 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, 2013, pp.1200-1204.

[10] Ahmad, X. Wang, R. Li and S. Rasheed, "Offline Urdu Nastaleeq optical character recognition based on stacked denoising autoencoder," in China Communications, vol. 14, no. 1, pp. 146 -157, Jan. 2017.

[11] Shetty, R. J., & Heraje, N. K. (2017). Recognition of Formatted Text using Machine Learning Technique.American Journal of Intelligent Systems, 7(3), 64-67.

[12] Allen TJ, Sherkat N, Whitrow RJ (1999) Holistic Word Case Recognition using a MultiLayer Perceptron Neural Network. IEE Colloquium on Document Image Processing and Multimedia.

[13] Bharath A, Madhvanath S (2007) Hidden Markov Models for Online Handwritten Tamil Word Recognition. IEEE ICDAR'2007, pp 23-26

[14] Chaudhuri BB, Pal U, Mitra M (2002) Automatic Recognition of Printed Oriya Script. Sadhana 27: 23-34.

[15] Cho W, Lee SW, Kim JH (1995) Modeling and Recognition of Cursive Words with Hidden Markov Models. Pattern Recognition 28(12): 1941-1953