A Project/Dissertation ETE REVIEW Report

on

HANDWRITTEN WORD IDENTIFIER

Submitted in partial fulfillment of the requirementfor the award of the degree of

BTech COMPUTER SCIENCE AND ENGINEERING

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision

Mr. S. KALIDASS

ASSISTANT PROFESSOR

Submitted By

JAI SHANKAR RAI 19SCSE1010297

TUSHAR CHOUDHARY 19SCSE1010058

SCHOOL OF COMPUTING SCIENCE AND

ENGINEERING DEPARTMENT OF COMPUTER

SCIENCE AND ENGINEERING

GALGOTIAS UNIVERSITY, GREATER NOIDA2021-22

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"HANDWRITTEN WORD IDENTIFER "** in partial fulfillment of the requirements for the award of the project review submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of month, Year to Month and Year, under the supervision of Name… Designation, Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Mr. S. KALIDASS

ASSISTANT PROFESSOR

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of JAI SHANKAR RAI 19SCSE1010297 & TUSHAR CHOUDHARY 19SCSE1010058 has been held on __________________ and his/her work is recommended for the award of

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December 2021
Place: Greater Noida

Acknowledgement

we would like to thank the our guide and reviewers for their valuable suggestions on our work.

Contents

Title		Page No.
Candidates Declaration		**I**
Acknowledgement		**II**
Abstract		**III**
Chapter 1	**Introduction**	**2**
Chapter 2	**Literature Survey**	**4**
Chapter 3	**Use case, Flow chart ,Data flow Diagram , Architectural Diagram**	**6**
Chapter 4	**Modules Description**	**8**
Chapter 5	**Result ,Conclusion and Future Scope**	**17**
	5.1 Conclusion	**18**
	5.2 Future Scope	**18**
	Reference	**19**

“ABSTRACT”

Handwritten Word Identifier (HWI) is one of the interesting fields in the image processing and words identifier application . Handwriting is the most natural mode of collecting, storing, and transmitting information which serves for communication of humans and machines. Many approaches are presented to recognize the handwritten documents or paper. These approaches focus on how we recognize hand written words and documents. Selecting the appropriate feature extraction methods and classifier is the most important thing in the handwritten word identifier. Hence achieving good recognition and better accuracy. Handwritten word identifier is the process of conversion of handwritten text into machine readable form which can be easily understandable to everyone. For handwritten characters there are difficulties like it differs from one writer to another, even when same person writes same character there is difference in shape, size and position of character, approaches focus on how we recognize hand written words and documents and it will also help in identify criminal activity written in bad writing and this aims to report the development of a Handwriting word identifier system that will be used to read students handwritten lectures notes.

“INTRODUCTION”

Handwriting Identifier is the ability of a machine to receive and interpret handwritten input from multiple sources like paper documents, photographs, touch screen devices etc. Identifying of handwritten and machine characters is an emerging area of research and finds extensive applications in banks, offices and industries, basically There are mainly 2 approaches in the document image processing. Online and offline approach, The main difference between these two categories is that the order of the strokes made by writer during writing is available to the online recognition system; whereas only the scanned version of the complete handwritten document is available to the offline handwritten word recognition system. Because of the availability of more relevant features, online recognition generally yields better results. Off-line handwritten text recognition is one of the most active areas of research in computer vision and it is inherently difficult because of the high variability in writing style, online Handwritten word Identifier the trajectories of pen tip movements are recorded and evaluated to identify intended information. This application is useful for identify all character given as in input image. Once input image of character is given to proposed system, then it will recognize input character which is given in image.Recognition and classification of characters are done by Neural Network.The main aim of this project is effectively recognize a particular character of type format using the Artificial Neural Network approach. An artificial neural Network as the backend is used for performing classification and recognition tasks. In the off-line recognition system, the neural networks have emerged as the fast and

reliable tools for classification towards achieving high recognition accuracy

A neural based off-line handwritten character recognition system, without feature extraction is proposed. The pre-processed image is segmented into individual characters. Each character is resized into 30x20 pixels and these pixels are used to train a feed forward back propagation neural network employed for performing classification and recognition tasks. Extensive simulation studies show that the recognition system provides good recognition accuracy.

Handwriting recognition has been one of the most fascinating and challenging research areas in field of image processing and pattern recognition in the recent years. It contributes immensely to the advancement of automation process and improves the interface between man and machine in numerous applications. Several research works have been focusing on new techniques and methods that would reduce the processing time while providing higher recognition accuracy.

An example of recognition of an abstract item is the recognition of solution to a problem when your eyes and ears are closed. However we are concerned only with recognition of concrete items. Recognition of concrete patterns by human beings may be considered as a physiological problem which involves a relationship between a person and a physical stimulus. When a person perceives a pattern, he makes an inductive inference and associates this perception with some general concepts or clues which he has derived from his past experience. Human recognition in reality is a question of estimating the relative odds that the input data can be associated with one of a set of known statistical populations which depend on our past experience and which form the clues and the a priori information for recognition. Thus, the problem of pattern recognition may be regarded as one of discriminating the input data not between individual patterns but between populations, via the search for features or invariant attributes among members of a population.

Applications of offline Handwritten word identifier:-

- Postal address identification.
- Writer’s handwriting identification.
- Bank check recognition.
- Signature Verification in banks.
- Historical documents.
- Identifying the words in inscriptions.

Banking sectors, Health care industries and many such organizations where handwritten documents are used regularly. HCR systems also find applications in newly emerging areas where handwriting data entry is required, such as development of electronic libraries, multimedia database etc.

Four methods of cursive handwritten word recognition.

- Holistic approach:-It is method in which entire word is recognized without splitting them by extracting features of entire word.
- Segmentation based Approach character are segmented from word.
- Character classification and segmentation are performed simultaneously by using appropriate learning method.
- Mixed Approach :- This system consist of combination of all the method.
- On-line character recognition. It is system in which recognition is performed when characters are under creation.
- Off-line character recognition. It is system in which first handwritten documents are generated, scanned, stored in computer and than they are recognized.

We present concise survey of available HCR for English language. HCR techniques are discussed with their strength and weaknesses. Different types of features are extracted and different types of classifiers are used to classify the input characters. The current study is focused on investigation of possible techniques to develop an offline HCR system for English language for both separate characters and cursive words.

One of the narrow fields of image processing is recognizing characters from an image, which is referred to as Optical Character Recognition (OCR). This method is about reading an image containing one or more characters, or reading a scanned text of typed or handwritten characters and be able to recognize them. A lot of research has been done in this field in order to find optimal techniques with a high accuracy and correctness. The most used algorithms that proved a very high performance are machine learning algorithms like Neural Networks and Support

Vector Machine. One of the main applications of OCR is recognizing handwritten characters. In this project, we will focus on building a mechanism that will recognize handwritten digits. We will be reading images containing handwritten digits extracted from the MNIST database and try to recognize which digit is represented by that image. For that we will use basic Image Correlation techniques, also referred to as Matrix Matching. This approach is based on matrices manipulations, as it reads the images as matrices in which each element is a pixel. It 2 overlaps the image with all the images in the reference set and find the correlation between them in order to be able to determine the digit it represents. The goal of this project is to apply and manipulate the basic image correlation techniques to build program and keep polishing and enhancing in order to investigate to which extent it can get improved. This would allow us to see how far we can go, in terms of accuracy and performance, but using just the very simple and basic techniques of matrix matching and without going into complicated methods like machine learning.

“Literature Reviews/Comparative study”

Handwriting character recognition is one of the research fields in computer vision, artificial intelligence, and pattern recognition, A computer application that performs handwriting recognition can be argued to have the ability to acquire and detecting characters in pictures, paper documents, and other sources and convert them into electronic format or machine- encoded form In this paper author has proposed system is to efficiently recognize the offline handwritten digits with a higher accuracy than previous works done. Also previous handwritten number recognition systems are based on only recognizing single digits and they are not capable of recognizing multiple numbers at one time.So the author has focused on efficiently performing segmentation for isolating the digits. [The system may obtain Handwriting sources from a piece of paper through optical scanning or intelligent word recognition. Also, the system may be designed to detect the movement of the pen tip on the screen. In other words, handwriting identification may involve a system detecting movements of a pen tip on the screen to get a clue of the characters being written, . Handwriting identification can be classified into two: offline recognition and online recognition. Offline handwriting recognition involved the extraction of text or characters from an image to have letter codes that can be used within a computer.

English is a West Germanic language originated in England. It is the official language of more than 60 countries. And it is the most commonly used language all over the world. English language has 26 letters. With 5 vowels and 21consonants.Upper case and lower case letters make total of 52 alphabet characters. Approximately 359 million people speak English as a first language. B. Gatos et .al [1] proposed an off-line cursive on a novel combination of two different modes of word image normalization and robust hybrid feature extraction. Here two types of features are combined. The first feature which creates the set of zones by dividing the image and calculates the density of the character pixels in each zone. Second feature calculates the area that is formed from the projections of the upper and lower profile of the word. Two classifiers are used here. Namely Minimum Distance Classifier and the Support Vector Machines (SVM). 80.76% accuracy is achieved using IAM database. Ankush Acharyya et .al [2] presented a holistic approach to recognize the offline handwritten words using Multi Layer Perceptron (MLP) classifier. Words are taken from the CMATERdb1.2.1 dataset. In this paper longest run features are used. These features are computed in four directions; row wise (east), column wise (north) and along the directions of two major diagonals (northeast and northwest). To get the more discriminating information of a particular word image, hierarchical partitioning is done till depth 5. Recognition rate achieved is 83%. Rodolfo Luna-Pérez, Pilar Gómez-Gil [3] described an unconstrained handwritten word recognition using a combination of neural networks. Author presented a novel method for classification of isolated handwritten words based on three components: a Self Organizing Map (SOM) for non-supervised classification of segments of a word, a function measuring probabilities of each segment belonging to a specific cluster and a Simple Recurrent Network (SRN) for temporal classification of a sequence of feature vectors obtained from segments forming the word. A Feed-Forward (FF) network, FF-SOM network are the classifiers used. 86.5% accuracy is achieved. IAM benchmark database is used for testing. Shaolei Feng et .al [4] reported Hidden Markov Model (HMM) for alphabet-soup word recognition. This approach first uses a joint boosting technique to detect potential International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org ICESMART-2015 Conference Proceedings Volume 3, Issue 19 Special Issue - 2015 2 characters –called as alphabet soup. In the second stage dynamic programming algorithm to recover the correct sequence of characters is described. A Hidden Markov Model is used to recognize a sequence of characters of fixed length given the character detection results. Here 85% words are recognized correctly.

Hindi is the national language of India. It is written in Devanagari script also called Nagari. Hindi is the native language of 280 million people. There are 11 vowels, 2 modifiers and 36 consonants makes total of 49 letters in Hindi language. Brijmohan Singh et .al [9] proposed a Curvelet Transform Based Approach to offline handwritten Devanagri word recognition. Principal Component Analysis (PCA) of the coefficients is used to reduce the size of feature vector to about 200 dimensions. For the recognition process Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) classifiers are used. K-NN produced better results than the SVM classifier and obtained 93.21% accuracy on Devanagari handwritten words. Vaibhav Dedhe, Sandeep Patil [10] reported a handwritten Devanagri special characters and word recognition using neural networks. The neural classifier consists of two hidden layers besides an input layer and an output layer. To extract the information of the boundary of a handwritten character, the eight-neighbor adjacent method is used. Proposed method has provided accuracy up-to 90% for special characters of Devanagari script. Naresh Kumar Garg et .al [11] described a offline handwritten Hindi text recognition using SVM method. The shape based features were extracted by applying many heuristics based on the shape of the character for each feature. Total 59 features are selected in the feature selection phase.89.6% recognition rate is achieved.

K. Gaurav and Bhatia P. K [2], proposed different prehandling systems being associated with the recognition of the characters. The procedure took a shot at the various types of pictures from a basic picture-based report to a hued and changed forces including foundation. Different systems of pre-handling and standardization like skew remedy, differentiate evacuation, commotion expulsion and numerous other upgrade procedures were recommended. They reached the decision that a solitary procedure can't be connected for preprocessing the picture. Yet additionally there were a few disparities that utilizing every one of these systems likewise can't give the best exactness comes about. Salvador España-Boquera [3], The analysts proposed the utilization of hybrid or half plus half concealed markov show (HMM) to perceive the handwritten content in disconnected mode. The optical model's basic part was prepared with markov chain procedure and a multilayer perceptron was likewise used to gauge the probabilities and perceive the disconnected handwritten numerals of six prominent Indian language, a changed quadratic classifier is utilized. A similar paper likewise manages perceiving the English letters in order. For both of these, a multilayer perceptron was utilized and Boundary following and Fourier descriptors were utilized for the component extraction. By examining the shape and looking at their highlights, the characters were identified. Also, to decide the quantity of concealed layers, back spread system was utilized. With this very calculation, a recognition rate of 94% have been accounted for with less preparing time.

Arabic is the native language of Arab countries. It has 290 million native speakers. And the official language of 27 countries. Arabic is the third most spoken language after English and Chinese. Arabic Abjad is the Arabic script used to write Arabic language. It is written from right to left. [5] The basic Arabic alphabet contains 28 letters.

Ahlam MAQQOR et .al [5] proposed a approach to cursive Arabic word recognition. The main objective of this system applies a multi-stream approach of two types of feature extraction methods. First one is based on local densities called as sliding window. and configurations of pixels and features a projection based on vertical, horizontal and diagonal 45 °, 135 ° - is the VH2D approach. By using multi-stream HMM 83.8% accuracy is obtained. Experiment is done on 200 Arabic words.

Ilya Zavorin [6] et.al described combining different classification approaches to Arabic handwritten word recognition. In this paper author spoke about the problem of offline Arabic handwriting recognition of pre-segmented words. Parts of Arabic Word (PAW) Segmenter, Ranking Lexicon Reducer, HMM Classifier are used here. Experiment is done on IFN/ENIT corpus of Tunisian village and town names.73% accuracy is achieved when combining the multiple classifiers. Alex Graves and Jurgen Schmidhuber [7] presented a offline Arabic handwritten word recognition using multidimensional recurrent neural networks. Author combined two methods in neural networks. Multidimensional recurrent neural networks and connectionist temporal classification. Instead of using single recurrent connection multidimensional recurrent neural networks are used. Because of this 91.7% accuracy is obtained. IFN/ENIT database of handwritten Arabic words is used for the experiment.

Volker Märgner et .al [8] described a offline handwritten word recognition of Arabic words using HMM method. Grey valued pixels of the normalized word image are used as features in the feature extraction steps. Sliding window and Karhunen-Loève Transform (KLT) are applied. Sequence of transformed feature vectors are used as the input to the HMM classifier. IFN/ENIT - Database is used in the experiment and got 89.77% recognition rate is achieved.

R. Bajaj, S. Chaudhari, L. Dey, et al [5], for grouping the Devanagari numerals, distinctive highlights like clear part, thickness and minute highlights were utilized. sAdditionally, to increase the recognition capacity, the paper proposes multi classifier unwavering quality for handwritten Devanagari numerals. Sandhya Arora in [6], In this paper specifically four highlights like shadow, histogram of chain code crossing point and horizontal line fitting highlights being portrayed. Among these highlights the shadow was registered all around for picture character, the rest three were processed by partitioning the character picture into the distinctive sections. In the one useful execution utilizing the dataset of 4900 examples demonstrated the exactness rate of 90.8 % for Devanagari handwritten characters.

J. Pradeep, E. Srinivasan and S. Himavathi (2012) [6] has designed Neural Network based recognition system. They used different neural network (NN) topologies- back propagation neural network, nearest neighbour network and radial basis function network for same training dataset. They compared the performance of each network and optimized the number of neurons in hidden layer which is not dependent on initial value and concluded that combination of standard feature extraction technique with feed forward back propagation.

M. Blumenstein, B. Verma and H. Basli (2003) [8]. This research describes neural network-based techniques for segmented character recognition. Two neural architectures along with two different feature extraction techniques were investigated. Directional and Transition features are used and compared by using Back-Propagation (BP) and Radial Basis Function (RBF) networks classifiers. The size of feature vector is 100 in case of transition feature and 81 for directional feature. Experiment was performed by using the CAS dataset, the BP (Back propagation) and RBF (Radial basis function) algorithm using two feature extraction techniques for both lower case and upper case characters, similarly for BAC database.

Rafael M. O. Cruz, George D. C. Cavalcanti and Tsang Ing Ren (2010) [11].In this paper recognization separate handwritten cursive characters is performed. Here different features are extracted among them two featuresmodified edge map and multiple zoning are proposed by authors. Total nine features are extracted and drawback of each feature each overcome by other. Each features are individually given as input to nine multilayer perceptron network and output of all this classifier are combined with each other by different rule like sum rule, product rule, max rule, mean rule etc among them trained MLP combiner gives maximum result. Among proposed features modified edge map feature gives highest result.

It involves obtaining digital data from a static representation of handwriting. A system is provided with a Handwriting document to read and convert the handwriting to a digital format. Online handwriting recognition, on the other hand, involved automatic detection or conversion of characters as they are written on the specialized screen.

NIST DATABASE:- Special Database 19 contains NIST's entire corpus of training materials for handprinted document and character recognition. It publishes Handprinted Sample Forms from 3600 writers, 810,000 character images isolated from their forms, ground truth classifications for those images, reference forms for further data collection, and software utilities for image management and handling.

The features of this database are:

- Final accumulation of NIST's handprinted sample data
- Full page HSF forms from 3600 writers
- Separate digit, upper and lower case, and free text fields
- Over 800,000 images with hand checked classifications

The database is NIST's largest and probably final release of images intended for handprint document processing and OCR research. The full page images are the default input to the NIST FORM-BASED HANDPRINT RECOGNITION SYSTEM, a public domain release of end-to-end recognition software.

Iam On-line Handwriting::- The IAM On-Line Handwriting Database (IAM-OnDB) contains forms of handwritten English text acquired on a whiteboard. It can be used to train and test handwritten text recognizers and to perform writer identification and verification experiments.

The database was first published in [LiBu05-03] at the ICDAR 2005. We use the database extensively in our own research, see publications for further details.

The database contains forms of unconstrained handwritten text, acquired with the E-Beam System. The collected data is stored in xml-format, including the writer-id, the transcription and the setting of the recording. For each writer the gender, the native language and some other facts which could be useful for the analysis are stored in the database.

The IAM Online Handwriting Database is structured as follows:

- 221 writers contributed samples of their handwriting
- more than 1'700 acquired forms
- 13'049 isolated and labeled text lines in on-line and off-line format
- 86'272 word instances from a 11'059 word dictionary

The Street View Text Dataset:- The Street View Text (SVT) dataset was harvested from Google Street View. Image text in this data exhibits high variability and often has low resolution. In dealing with outdoor street level imagery, we note two characteristics. (1) Image text often comes from business signage and (2) business names are easily available through geographic business searches. These factors make the SVT set uniquely suited for word spotting in the wild: given a street view image, the goal is to identify words from nearby businesses. More details about the data set can be found in our paper, Word Spotting in the Wild. For our up-to-date benchmarks on this data, see our paper, End-to-end Scene Text Recognition. This dataset only has word-level annotations (no character bounding boxes) and should be used for (A) cropped lexicon-driven word recognition and (B) full image lexicon-driven word detection and recognition. If you need character training data then you should look into the Chars74K and ICDAR datasets **.**

Scene Text: Contains 3000 images captured in different environments, including outdoors and indoors scenes under different lighting conditions (clear day, night, strong artificial lights, etc).

MSRA Text Detection 500 Database (MSRA-TD500)

The MSRA Text Detection 500 Database (MSRA-TD500) is collected and released publicly as a benchmark to evaluate text detection algorithms, for the purpose of tracking the recent progresses in the field of text detection in natural images, especially the advances in detecting texts of arbitrary orientations.

The MSRA Text Detection 500 Database (MSRA-TD500) contains 500 natural images, which are taken from indoor (office and mall) and outdoor (street) scenes using a pocket camera. The indoor images are mainly signs, doorplates and caution plates while the outdoor images are mostly guide boards and billboards in complex background. The resolutions of the images vary from 1296x864 to 1920x1280.

The dataset is challenging because of both the diversity of the texts and the complexity of the background in the images. The text may be in different languages (Chinese, English or mixture of both), fonts, sizes, colors and orientations. The background may contain vegetation (e.g. trees and bushes) and repeated patterns (e.g. windows and bricks), which are not so distinguishable from text.

The dataset is divided into two parts: training set and test set. The training set contains 300 images randomly selected from the original dataset and the remaining 200 images constitute the test set. All the images in this dataset are fully annotated. The basic unit in this dataset is text line (see Figure 1) rather than word, which is used in the ICDAR datasets, because it is hard to partition Chinese text lines into individual words based on their spacing; even for English text lines, it is non-trivial to perform word partition without high level information.

HANA: A HAndwritten Name Database for Offline Handwritten Text Recognition:- Methods for linking individuals across historical data sets, typically in combination with AI based transcription models, are developing rapidly. Probably the single most important identifier for linking is personal names. However, personal names are prone to enumeration and transcription errors and although modern linking methods are designed to handle such challenges these sources of errors are critical and should be minimized. For this purpose, improved transcription methods and large-scale databases are crucial components. This paper describes and provides documentation for HANA, a newly constructed large-scale database which consists of more than 1.1 million images of handwritten word-groups. The database is a collection of personal names, containing more than 105 thousand unique names with a total of more than 3.3 million examples. In addition, we present benchmark results for deep learning models that automatically can transcribe the personal names from the scanned documents. Focusing mainly on personal names, due to its vital role in linking, we hope to foster more sophisticated, accurate, and robust models for handwritten text recognition through making more challenging large-scale databases publicly available. This paper describes the data source, the collection process, and the image-processing procedures and methods that are involved in extracting the handwritten personal names and handwritten text in general from the forms.

MNIST Database: A subset of the original NIST data, has a training set of 60,000 examples of handwritten digits, The MNIST database, which stands for the Modified National Institute of Standards and Technology database, is a very large dataset containing several thousands of handwritten digits. This dataset was created by mixing different sets inside the original National Institute of Standards and Technology (NIST) sets, so as to have a training set containing several types and shapes of handwritten digits, as the NIST set was divided into those written by high school students and others written by the Census Bureau workers [14]. The MNIST dataset 8 has been the target of so many research done in recognizing handwritten digits. This allowed the development and improvements of many different algorithms with a very high performance, such as machine learning classifiers.

EMNIST dataset is extended by adding some more characters from Tamil language.

KAGGLE DATASET:- For recognising handwritten forms, the very first step was to gather data in a considerable amount for training. Which I struggled to collect for weeks. The dataset contains 26 folders (A-Z) containing handwritten images in size 28*28 pixels, each alphabet in the image is centre fitted to 20*20 pixel box.Each image is stored as Gray-levelKernel CSV*To*Images contains script to convert .CSV file to actual images in .png format in structured folder.

UJI Pen Characters Data Set: e a character database by collecting samples from 11 writers. Each writer contributed with letters (lower and uppercase), digits, and other characters (Spanish diacritics and punctuation marks) that we have not employed in our experiments and are not included in this database version. Two samples have been collected for each pair writer/character, so the total number of samples in this database version is 1364:
11 writers x 2 repetitions x (2x26 letters + 10 digits)
The proposed task is a writer-independent one consisting of 11 leaving-one-writer-out tests, so the effective training set size (for each one of the 1364 test samples) is 1240:
10 writers x 2 repetitions x (2x26 letters + 10 digits)
Moreover, this classification task is a 35-class one because we have not considered a different class for each different character: each one of the 26 letters is considered as a case-independent class, there are 9 additional clases for non-zero digits, and the zero is included in the same class as o's.

Paper Name	AUTHOR	FEATURE USED	CLASSIFIERS	DATASET	ACCURACY
“Efficient Offline Cursive Handwriting Word Recognition”	] B. Gatos, I. Pratikakis, A.L. Kesidis, S.J. Perantonis	Zones and upper and lower profile of the word	Minimum Distance Classifier and the SVM	IAM	80.76%
“Handwritten Word Recognition Using MLP based Classifier: A Holistic Approach	Ankush Acharyya, Sandip Rakshit, Ram Sarkar, Subhadip Basu, Mita Nasipuri,	Longest run features	MLP classifier.	CMATERdb1.2.1	83%
“Unconstrained Handwritten Word	Rodolfo Luna-Perez, Pilar Gomez-Gil	Feature extractor based on non-	Feed-forward (FF) network, FF-SOM	IAM	86.5%

Recognition Using a Combination of Neural Networks”,		supervised clustering	network		
A Hidden Markov Model for Alphabet-Soup Word Recognition”	Shaolei Feng Nicholas R. Howe R. Manmatha	Joint boosting technique	Hidden markov model	Mnist	85%
A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters”	M. Blumenstein, B. Verma, H. Basli (2003)	Directional and Transition features	BP and RBF networks	BAC and CAS	85%
”Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten”	J. Pradeep, E. Srinivasan S. Himavathi (2012)	character resized into 30X20 pixels taken as feature	Neural Network	Kaggle	94.15%
An Ensemble classifier for offline Cursive character recognition using multiple features Extraction technique”	Rafael M. O. Cruz, George D. C. Cavalcanti and Tsang Ing Ren (2010)	Modified Edge Maps and Multi zoning.	MLP	C-Cube Database	91%
“Writer Identification by Texture Analysis Based on Kannada Handwriting”	B.V.Dhandra et .al	Discrete Cosine Transform, Gabor filtering and gray level cooccurrence matrix	K-NN classifier	Own dataset	88.5%

What is Neural Network ?

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 large no. of highly interconnected processing element (neurons) working in union to solve specific problems. ANN‘s like peopling, learning by example. An ANN is configured for a specific application. such as pattern recognition or data classification, through a learning process .Learning in a Biological system involves adjustments to the synaptic connections that exist between the neuron.

Why use Neural Network ?

Neural network with their remarkable ability to derive meaning from complicated or imprecise data can be use to extract pattern and detect trend that are too complex to be noticed by either human or other computer techniques. A trained neural network can be thought of as an “expert” in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer “what if” questions. Other Advantages Include: • Adaptive Learning: An ability to learn how to do tasks based on the data given for training or initial experience. • Self-Organization:An ANN can create its own organization or representation of the information it receives during learning time. • Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability. • Fault Tolerance Via Redundant Information coding: partial destruction of network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.

Implementation of HCR :- HCR works in stages as preprocessing, segmentation, feature extraction and recognition using neural network. Preprocessing includes series of operations to be carried out on document image to make it ready for segmentation. During segmentation the document image is segmented into individual character or numeric image then feature extraction technique is applied on character image. Finally feature vector is presented to the selected algorithm for recognition.Here this extracted features are provided to NN for recognition of character.

Recognition of Handwritten:- Hindi Characters using Backpropagation Neural Network Automatic recognition of handwritten characters is a difficult task because characters are written in various curved and cursive ways, so they could be of different sizes, orientation, thickness, format and dimension. An offline handwritten Hindi character recognition system using neural network is presented in this paper. Neural networks are good at recognizing handwritten characters as these networks are insensitive to the missing data. The paper proposes the approach to recognize Hindi characters in four stages 1) Scanning, 2) Preprocessing, 3) Feature Extraction and, 4) Recognition. Preprocessing includes noise reduction, binarization, normalization and thinning. Feature extraction includes extracting some useful information out of the thinned image in the form of a feature vector. The feature vector comprises of pixels values of normalized character image. A Back propagation neural network is used for classification. Experimental result shows that this approach provides better results as compared to other techniques in terms of recognition accuracy, training time and classification time. The average accuracy of recognition of the system is 93 %

Intelligent Systems for Off-Line Handwritten Character Recognition: A Review Handwritten character recognition is always a frontier area of research in the field of pattern recognition and image processing and there is a large demand for Optical Character 4 Recognition on hand written documents. This paper provides a comprehensive review of existing works in handwritten character recognition based on soft computing technique during the past decade.

An Overview of Character Recognition Focused on Off-Line Handwriting Character recognition (CR)

has been extensively studied in the last half century and progressed to a level sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present CRmethodologies and creates an increasing demand on many emerging application domains, which require more advanced methodologies

Recognition of Handwritten Devnagari Characters through Segmentation and Artificial neural networks Handwritten character recognition is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touchscreens and other devices.Handwritten Marathi Characters are more complex for recognition than corresponding English characters due to many possible variations in order, number, direction and shape of the constituent strokes. The main purpose of this paper is to introduce 5 a new method for recognition of offline handwritten devnagari characters using segmentation and Artificial neural networks. The whole process of recognition includes two phasessegmentation of characters into line, word and characters and then recognition through feedforward neural network

Image preprocessing for optical character recognition using neural networks :- Primary task of this master's thesis is to create a theoretical and practical basis of preprocessing of printed text for optical character recognition using forward-feed neuralnetworks. Demonstration application was created and its parameters were set according to results of realized experiments.

Recognition for Handwritten English Letters: A Review Character recognition is one of the most interesting and challenging research areas in the field of Image processing. English character recognition has been extensively studied in the last half century. Nowadays different methodologies are in widespread use for character recognition. Document verification, digital library, reading bank deposit slips, reading postal addresses, extracting information from cheques, data entry, applications for credit cards, health insurance, loans, tax forms etc. are application areas of digital document processing. This paper gives an overview of research work carried out for recognition of hand written English letters. In Hand written text there is no constraint on the writing style. Hand written letters are difficult to recognize due to diverse human handwriting style, variation in angle, size and shape of letters. Various approaches of hand written character recognition are discussed here along with their performance.

Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Network:- An off-line handwritten alphabetical character recognition system using multi layer feed forward neural network is described in the paper. A new method, called, diagonal based feature extraction is introduced for extracting the features of the handwritten alphabets.

Fifty data sets, each containing 26 alphabets written by various people, are used for training the neural network and 570 different handwritten alphabetical characters are used for testing. The proposed recognition system performs quite well yielding higher levels of recognition 6 accuracy compared to the systems employing the conventional horizontal and vertical methods of feature extraction. This system will be suitable for converting handwritten documents into structural text form and recognizing handwritten names.

Fuzzy Based Handwritten Character Recognition System:- This paper presents a fuzzy approach to recognize characters. Fuzzy sets and fuzzy logic are used as bases for representation of fuzzy character and for recognition. This paper describes a fuzzy based algorithm which first segments the character and then using fuzzy system gives the possible characters that match the given input and then using defuzzication system finally recognizes the character.

Machine Learning Classifiers

A classifier in machine learning is an algorithm that automatically orders or categorizes data into one or more of a set of "classes." One of the most common examples is an email classifier that scans emails to filter them by class label: Spam or Not Spam.Machine learning algorithms are helpful to automate tasks that previously had to be done manually. They can save huge amounts of time and money and make businesses more efficient. Machine learning classifiers are used to automatically analyze customer comments (like the above) from social media, emails, online reviews, etc., to find out what customers are saying about your brand.Other text analysis techniques, like topic classification, can automatically sort through customer service tickets or NPS surveys, categorize them by topic (Pricing, Features, Support, etc.), and route them to the correct department or employee.

- **Classifier:** An algorithm that maps the input data to a specific category.
- Classification model**:** A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.
- **Feature:** A feature is an individual measurable property of a phenomenon being observed.
- Binary Classification**:** Classification task with two possible outcomes. Eg: Gender classification (Male / Female)
- Multi-class classification**:** Classification with more than two classes. In multi class classification each sample is assigned to one and only one target label. Eg: An animal can be cat or dog but not both at the same time
- Multi-label classification**:** Classification task where each sample is mapped to a set of target labels (more than one class). Eg: A news article can be about sports, a person, and location at the same time.

Artificial Intelligence :- The idea of reading Handwriting characters, digits, and words by computer systems can be argued to be an imitation of a human being. In other words, such a system can be argued that they use artificial intelligence to read handwriting from images or any Handwriting source, Artificial intelligence refers to intelligence that is demonstrated by machines The term is used to describe computer or machines that can mimic "cognitive" functions that are associated with the human mind. Artificial intelligence allows the machine tolearn from experience, adjust to new data (inputs), and perform tasks that can be performed byhumans.

Machine Learning :-Machine learning technology is inspired by psychology and biology that focus on learning from a set of data. The central assumption is that machines can learn to perform given tasks by learning from data. A machine learning model is provided with training data that is specific to the given problem domain and the solution to each instance of the problem. That way, the model learns how to solve certain problems based on learning.

Artificial Neural Network (ANN) :-Artificial Neural Network (ANN) refers to information processing paradigm or computing systems that are inspired by biological neural networks that constitute the human brain . The systems are not identical to the biological neural systems, but they are designed to process information the same way the human brain and animal brain process information . The networks are composed of many interconnected neurons working in unison to achieve specific goals . Just like the human brain, ANN learns from example. Hence, an ANN can be configured for an application, such as data classification or character recognition through the learning process. The learning process involves adjusting the system to a connection . The artificial neural network comprises a network of multiple simple processors, each with a small amount of local memory.

Deep Neural Network:- The neural network has layers of units where each layer takes some value from the previous layer. That way, systems that are based on neural networks can compute inputs to get the needed output[4] . The same way neurons pass signals around the brain, and values are passed from one unit in an artificial neural network to another to perform the required computation and get new value as output . Neural Processing Unit (NPU) architectures dedicated to energy-efficient DNN acceleration became essential. Despite the fact that training phase of DNN requires precise number representations, many researchers proved that utilizing smaller bit-precision is enough for inference with low-power consumption. This led hardware architects to investigate energy-efficient NPU architectures with diverse HW-SW co-optimization schemes for inference. This chapter provides a review of several design examples of latest NPU architecture for DNN, mainly about inference engines The united are layers, forming a system that starts from the layers used for imputing to layer that is used to provide the output. The layers that are found between the input and output layers are called the hidden layer. The hidden layers refer to a deep neural network that is used for computation of the values inputted in the input layer, h enough training data, the deep neural network can be able to perform any function that a neural network is supposed to do. It is only possible if the neural network has enough hidden layers, although the smaller deep neural network is more computationally efficient than a more extensive deep neural network.

Convolutional Neural Networks(CNN):-CNN stands for

Convolutional Neural Networks that are used to extract the features of the images using several layers of filters[10]. Convolutional networks are a specialized type of neural networks that use convolution in place of general matrix multiplication in at least one of their layers. The convolution layers are generally followed by maxpool layers that are used to reduce the number of features extracted and ultimately the output of the maxpool and layers and convolution layers are flattened into a vector of single dimension andare given as an input to the Dense layer (The fully connected network).

convolutional network were inspired by biological processes[9][10][11][12] in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field.CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns to optimize the filters (or kernels) through automated learning, whereas in traditional algorithms these filters are hand-engineered. This independence from prior knowledge and human intervention in feature extraction is a major advantage.

Convolutional neural network is used as a classifier for classifying the handwritten character from the input image. PAPER NAME FEATURE USED CLASSIFIERS ACCURACY "Efficient Offline Cursive Handwriting Word Recognition[5]" Zones and upper and lower profile of the word Minimum Distance Classifier and the SVM 80.76% "Handwritten Word Recognition Using MLP based Classifier: A Holistic Approach[6] Longest run features MLP classifier. 83% "Unconstrained Handwritten Word Recognition Using a Combination of Neural Networks[7]", Feature extractor based on non□supervised clustering Feed-forward (FF) network, FF-SOM network 86.5% A Hidden Markov Model for Alphabet□Soup Word Recognition[8]" Joint boosting technique Hidden markov model 85% Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks[9] The hierarchical structure Multidimensional recurrent neural networks 91% A CNN consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers and normalization layers[15]. A CNN consists of three major components which are convolutional layer, pooling layer and output layer. The activation function that is commonly used with CNN is ReLU which stands for Rectified Linear Unit. Convolution layer will compute the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume. The pooling layer is a form of nonlinear down sampling. Max pooling is the most common which partition the input image into a set of non-overlapping rectangles and, for each such sub-region, outputs the maximum. ReLU applies the non-saturating activation function . It increases the nonlinear properties of the decision function and of the overall network without affecting the receptive fields of the convolution layer. A rectified linear unit has output 0 if the input is less than 0, and raw output otherwise. Its value is obtained based on the formula which is as follows: f(x)= max(x,0) The softmax function is often used in the final layer of a neural network-based classifier. The softmax function squashes the outputs of each unit to be between 0 and 1, just like a sigmoid function. But it also divides each output such that the total sum of the outputs is equal to 1. The output of the softmax function is equivalent to a categorical probability distribution. Thus, softmax function calculates the probabilities distribution of the event over 'n' different events

Hidden Markov Models (HMM):- Hidden Markov Model (HMM) has been used in many handwriting recognition systems as a primary modeling component. HMM is a statistical Markov model that is used in a system that is supposed to assume the Markov process . It can be considered as the most straightforward dynamic Bayesian network. Hidden Markov Models are class pf probabilistic graphical models used for predicting a sequence of hidden variables from a set of observed variables More specifically, you only know observational data and not information about the states. In other words, there's a specific type of model that produces the data (a Markov Model) but you don't know what processes are producing it. You basically use your knowledge of Markov Models to make an educated guess about the model's structure.

METHODOLOGY

The current OCR system will consist of five phases. The phases are image acquisition and digitization, preprocessing, segmentation, feature extraction, and recognition.

A system is developed to recognize handwritten English characters using CNN, for a subset of the English characters.

The collected databases are divided into two parts Training data and testing data. Training data are used to train the system and this trained system are than used to recognize test data, The block diagram of the handwritten character recognition system is shown in figure 1

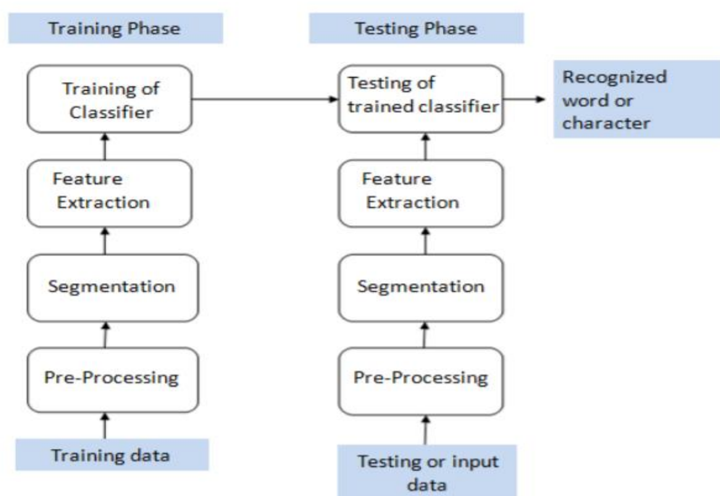


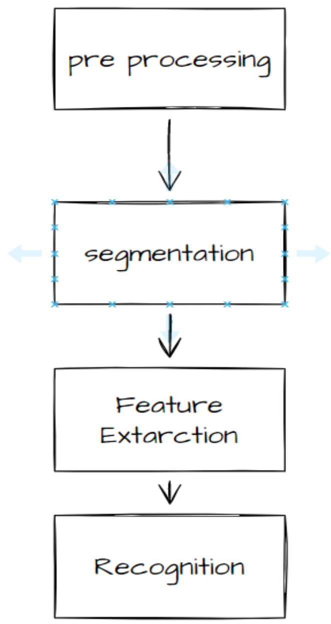
Fig 1: Block diagram of HCR System

1. Preprocessing

2 Segmentation .

3 Feature Extraction

4 Recognition



1. PREPROCESSING:-Preprocessing is concerned mainly with the reduction of these kinds of noise and variability in the input Preprocessing is essential for developing data that that are easy for optical character recognition systems. [10]. The number and type of preprocessing algorithms employs on the scanned image depend on many factors such as paper quality, resolution of the scanned image, the amount of skew in the image and the layout of the text. Various preprocessing operations are performed prior to recognition to enhance the quality of the input image, .the digitized image is pre-processed to remove noise, and then it is checked for skewing. Preprocessing is essential for developing data that that are easy for optical character recognition systems. The main objective of pre-processing is to remove the background noise, enhance the region of interest in the image, and make a clear difference between foreground and background.

1) Image enhancement techniques: To modify attributes of the image to make it more suitable and to improve the quality of the image by reducing noise, increasing contrast, image blurring, and providing more details. Hence, to

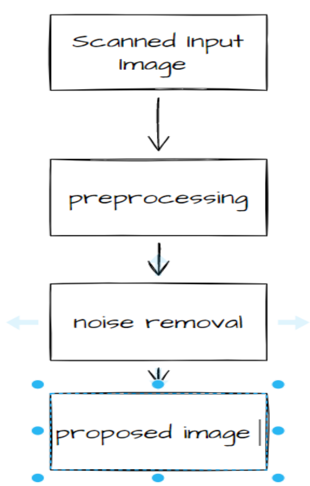
process an image so that result is more suitable than the original image and providing better input for automated image processing techniques.

2) Noise removal: Addictive noises of different types can contaminate images. Hence there is a need to remove noise to improve the quality of the image.

3) Binarization: This method is used to transform the grayscale image and converting it to black and white, substantially reducing the information contained within the image from different shapes of gray into a binary image.

4) Normalization: This process in image processing that changes the range of pixel intensity values. Its common purpose of converting an input image into a range of pixel values that are more familiar to the senses. Normalization involves converting images into a standard size.

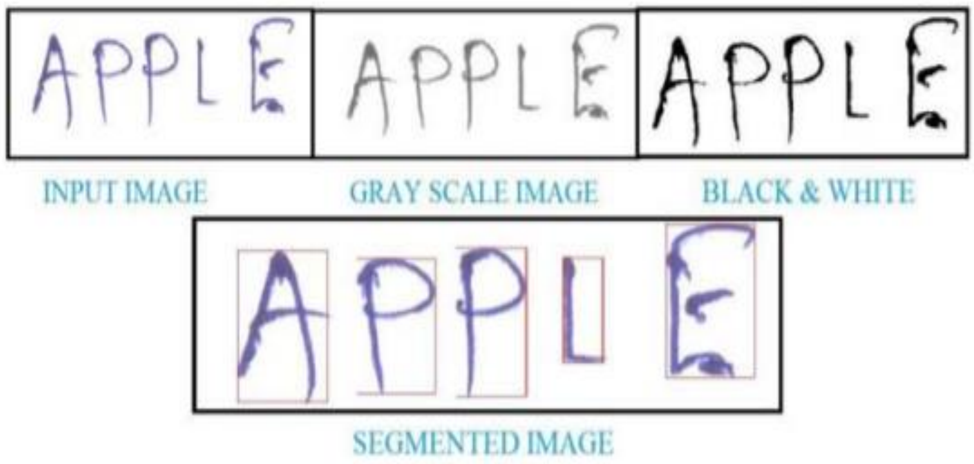
5) Skew correction, thinning: This is one of the first operations to be applied to scanned documents when converting data to digital format. This process helps to get a single-pixel width to allow easy character recognition.



2.Segmentation:- Segmentation can be argued to be the most critical process in character recognition techniques. Segmentation of images is done in the testing stage only. It checks for any error point inclusion by checking all points against the average distance between segmentation points incomplete image[12]. The process involves separating individual characters from an image. The process results in multiple segments of the image known as super pixels. The main aim of segmentation is to simplify the representation of an image into something that can be analyzed easily. Hence it has a positive impact on the recognition rate of the script, . It checks for any error point inclusion by checking all points against the average distance between segmentation points incomplete image. The process involves separating individual characters from an image,. The process results in multiple segments of the image known as super pixels. The main aim of segmentation is to simplify the representation of an image into something that can be analyzed easily. Hence it has a positive impact on the recognition rate of the script line segmentation which is separation of line from paragraph, the pre-processed input image is segmented into isolated characters by assigning a number to each character using a labeling process. This labeling provides information about number of characters in the image. Each individual character is uniformly resized into 30X20 pixels.

- Word segmentation which is separation of word from line.
- Character segmentation which is separation of character from words, Character segmentation is performed if segmentation based method is adopted for cursive word recognition, for holistic method character segmentation is not performed.

After the image is cleaned up and becomes a binary image which contains only the text, the binary image is then saved and the memory is cleaned up.This step is very important to increase the speed of the system.After the following steps should be done. • Divide the text into rows • Divide the rows intowords • Divide the word into letters



3:- Feature Extraction:-In this phase, features of the image are extracted and are defined based on the following attributes: height of the character, numbers of horizontal lines, widths of the character, number of circles, pixels, position of different features and number of vertically oriented arcs, to mention a few[14]. Feature Extraction is the problem of extracting from the pre processed data, the information, which is most relevant for classification purposes, in the sense of minimizing the within-class pattern variability, while enhancing the between-class pattern variability. There by flattening the array into a vector of 28*28 = 784 numbers. Thus, the image now converges to a minimal bunch of arrays in a 784-cell dimension of a highly efficient structure. The image now becomes a tensor of n dimensional array, features of the image are extracted and are defined based on the following attributes: height of the character, numbers of horizontal lines, widths of the character, number of circles, pixels, position of different features and number of vertically oriented arcs, to mention a few, The main objective of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements. Feature extraction is the heart of any pattern recognition application. Feature extraction techniques like

Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), Scale Invariant Feature Extraction (SIFT), zoning, Gradient based features, Histogram might be applied to extract the features of individual characters. f these characteristics would be the curvatures, the holes, the edges, etc. In the case of digits recognition, these features could be the holes inside the digits (for example for the eight, the six, and maybe the two as well) as well as the angles between some straight lines (for example in the one, the four, and the 6 seven). Whenever an unknown image is to be recognized, its features are compared to these so that it can be classified.

These features are used to train the system. The features extracted are given as input to the classification stage and the output of this stage is a recognised character. The selection of the combination of feature-classifier contributes to the performance of the system. Several research works has been focussing toward evolving such methods to reduce the processing time and providing higher recognition . Feature extraction based on three types of feature:

1.) Statistical Feature These features are derived from statistical distribution points. They provide high speed and low complexity and take care of style variation. Zoning, characteristic loci, crossing and distance are the main statistical features.

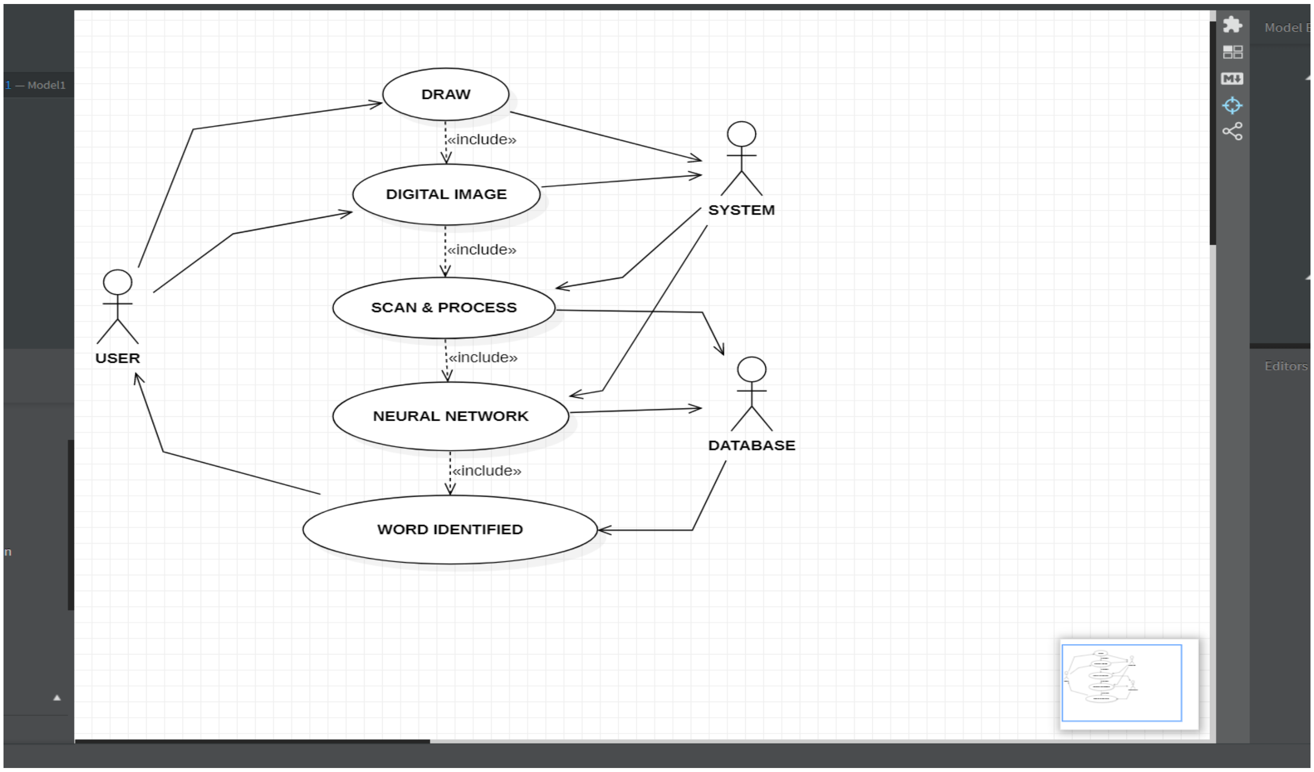
2.) Structural Features Structural features are based on topological and geometrical properties of character, such as, aspect ratio, cross point, loops, branch points, strokes and thire directions, inflection between two points, horizontal curve on top or bottom, etc. The representation of this type may also encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object.

3.) Global transformation and series expansion A continuous signal generally contains more information than needs to be represented for the purpose of classification[4]. This may be true for discrete approximations of continuous signals as well. One way to represent a signal is by a linear combination of a series of simpler well-defined functions. The coefficients of the linear combination give a compact encoding known as transformation or/and series expansion. Deformations like translation and rotations are invariant under global transformation and series expansion. Gabor transformation, This features stores information contained in whole image in few coefficients, thus it performs energy compactness. Various types of global transformation based features are: Discrete Fourier Transform, Discrete Cosine Transform, Discrete Wavelet Transform etc. Fourier transformation and wavelet transformation are common transform and series expansion method used in character recognition method.

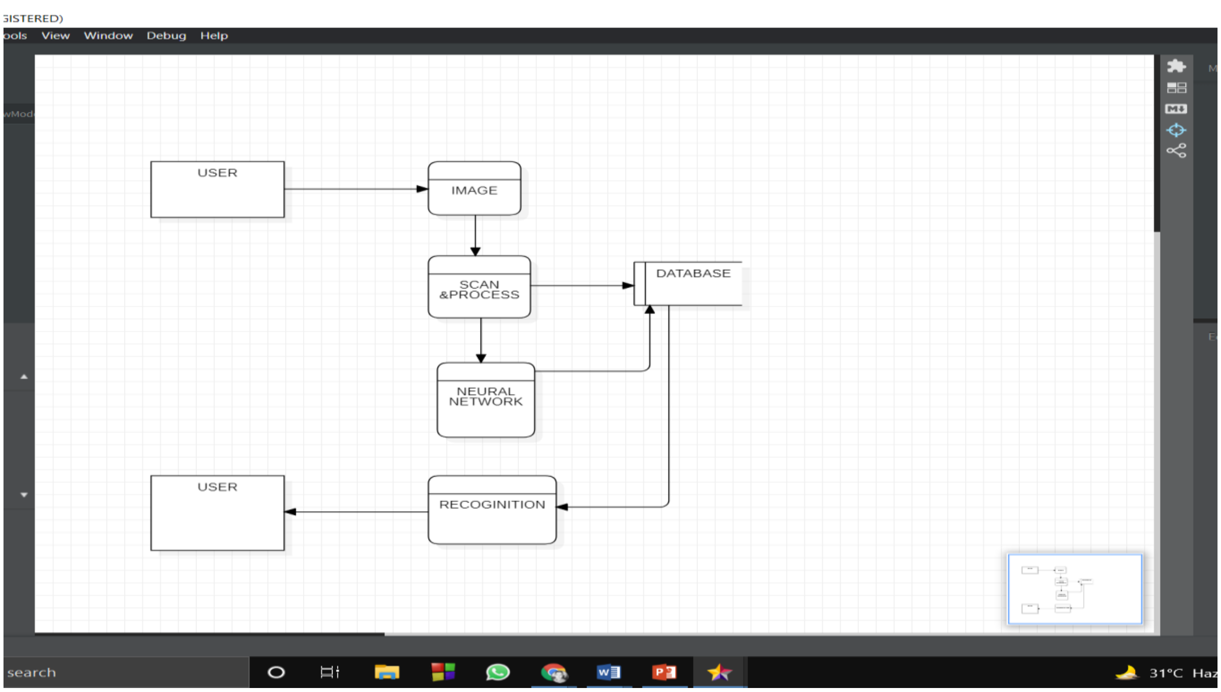
4. Recognition: The neural network is used for classification and recognition of the characters from the image. The most neural networks that are used by optical character recognition systems are the multiplayer perception (MLP) and Korhonen's Self Organizing Map, Extracted features in the feature extraction steps are used to classify the images by assigning labels to these features. The data set collected is divided into training data set and testing data set. Bayesian classifier, Binary tree classifier, Nearest Neighbor classifier, Neural networks, Hidden Markov Model (HMM) and Support Vector Machines (SVM) are some of the classifiers that are used in this stage. the decision making part of a recognition system and it uses the features extracted in the previous stage. The feature vector is denoted as X where X = (f1, f2,....., fd) where f denotes features and d is the no. features extracted from character. Based on the comparison of feature vector characters are efficiently classified into appropriate class and recognized. Classifiers are based on two types of learning methods. • Supervised

learning In supervised learning training data with correct detail of class is applied to train a model. This model is used to test data for proper classification. Training data includes both the input and the desired results. The model undergoes learning process and based on this learning it classifies test data. For example: SVM, HMM etc. • Unsupervised learning In unsupervised learning model is not provided with training data. It does not require learning. The model classifies test data based on statistical properties and by their spatial grouping and considering their nearest neighbour. For example: Clustering, k means etc

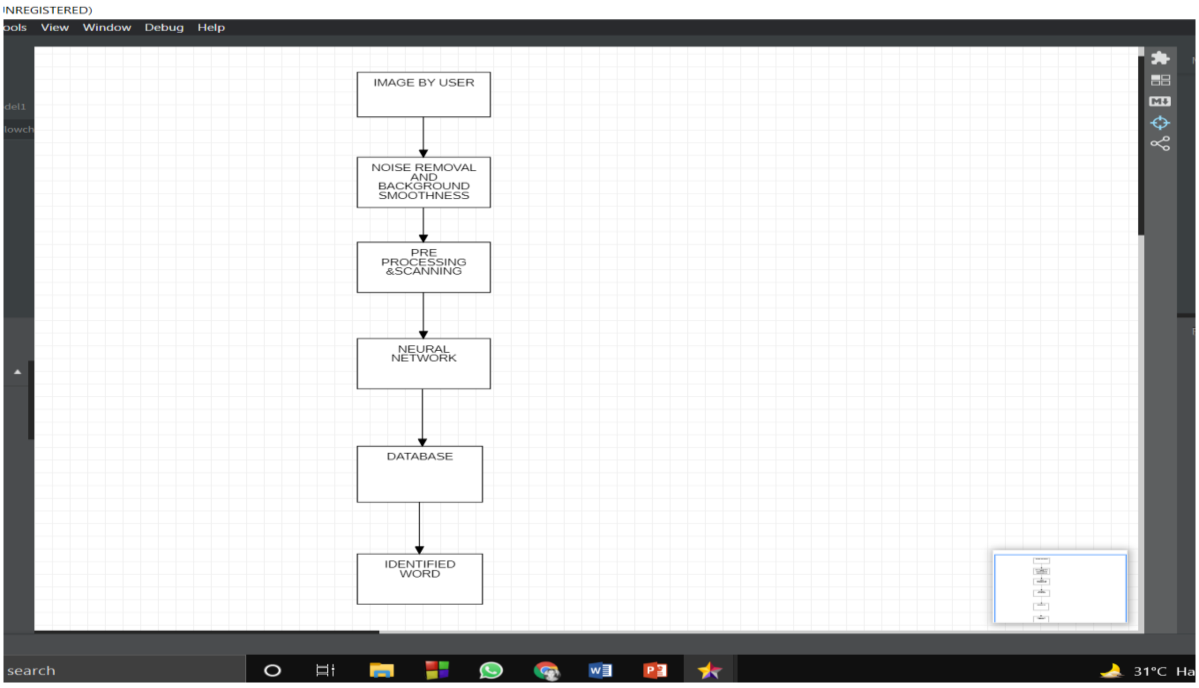
USECASE DIAGRAM



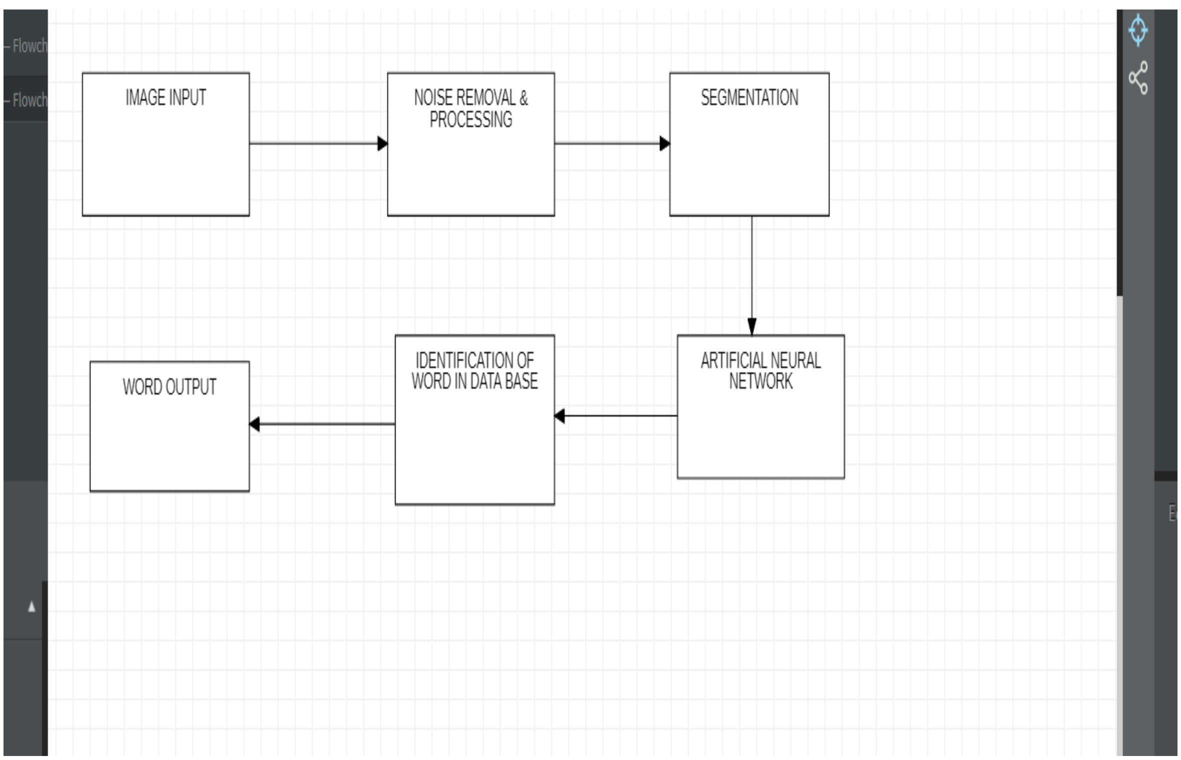
DATA FLOW DIAGRAM



FLOW CHART



ARCHITECTURAL DIAGRAM



Modules Description

Project Prerequisites

Below are the prerequisites for this project:

1. Python (3.7.4 used)
2. IDE (Jupyter used)

Required frameworks are

1. Numpy
2. cv2 (openCV)
3. Keras
4. Tensorflow (Keras uses TensorFlow in backend and for some image preprocessing)
5. Matplotlib Pandas)

Download Dataset

The dataset for this project contains 372450 images of alphabets of 28×28, all present in the form of a CSV file

Kaggle Dataset

A-Z Handwritten Alphabets in .csv format

The dataset contains 26 folders (A-Z) containing handwritten images in size 28*28 pixels, each alphabet in the image is centre fitted to 20*20 pixel box.

Each image is stored as Gray-level

Kernel CSV*To* Images contains script to convert .CSV file to actual images in .png format in structured folder.

Important library :-

- import matplotlib.pyplot as plt:- Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB.

- import cv2:- OpenCV is a great tool for image processing and performing computer vision tasks. It is an open-source library that can be used to perform tasks like face detection, objection tracking, landmark detection, and much more. It supports multiple languages including python, java C++

- import numpy as np:- NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices.

- **keras:-** Keras is a powerful and easy-to-use free open source Python library **for developing and evaluating deep learning models**. It wraps the efficient numerical computation libraries Theano and TensorFlow and allows you to define and train neural network models in just a few lines of code.
- from keras.models import Sequential:- The core idea of ***Sequential API*** is simply arranging the Keras layers in a sequential order and so, it is called *Sequential API*.
- from keras.layers import Dense, Flatten, Conv2D, MaxPool2D, Dropout
- from tensorflow.keras.optimizers import Adam
- from tensorflow.keras.optimizers import SGD
- from keras.callbacks import ReduceLROnPlateau, EarlyStopping
- f from tensorflow.keras.utils import to_categorical
- import pandas as pd
- import numpy as np
- from sklearn.model_selection import train_test_split
- from sklearn.utils import shuffle

Read the data:

```
data = pd.read_csv(r"D:\a-z alphabets\A_Z Handwritten
Data.csv").astype('float32')
print(data.head(10))
```

Now we are reading the dataset using the **pd.read_csv()** and printing the first 10 images using **data.head(10)**

```
In [2]: data = pd.read_csv(r"D:\archive\A_Z Handwritten Data\A_Z Handwritten Data.csv").astype('float32')
        print(data.head(10))

     0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9  ...  0.639  0.640  0.641  \
0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
1  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
2  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
3  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
4  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
5  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
6  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
7  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
8  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0
9  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  ...    0.0    0.0    0.0

   0.642  0.643  0.644  0.645  0.646  0.647  0.648
0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
1    0.0    0.0    0.0    0.0    0.0    0.0    0.0
2    0.0    0.0    0.0    0.0    0.0    0.0    0.0
3    0.0    0.0    0.0    0.0    0.0    0.0    0.0
4    0.0    0.0    0.0    0.0    0.0    0.0    0.0
5    0.0    0.0    0.0    0.0    0.0    0.0    0.0
6    0.0    0.0    0.0    0.0    0.0    0.0    0.0
7    0.0    0.0    0.0    0.0    0.0    0.0    0.0
8    0.0    0.0    0.0    0.0    0.0    0.0    0.0
9    0.0    0.0    0.0    0.0    0.0    0.0    0.0

[10 rows x 785 columns]

In [3]: X = data.drop('0',axis = 1)
```

Split data into images and their labels:

```
X = data.drop('0',axis = 1)
y = data['0']
```

Splitting the data read into the images & their corresponding labels. The ‘0’ contains the labels, & so we drop the ‘0’ column from the data dataframe read & use it in the y to form the labels.

Reshaping the data in the csv file so that it can be displayed as an image

```
train_x, test_x, train_y, test_y = train_test_split(X, y, test_size = 0.2)

train_x = np.reshape(train_x.values, (train_x.shape[0], 28,28))
test_x = np.reshape(test_x.values, (test_x.shape[0], 28,28))

print("Train data shape: ", train_x.shape)
print("Test data shape: ", test_x.shape)
```

the above segment, we are splitting the data into training & testing dataset using train_test_split().

Also, we are reshaping the train & test image data so that they can be displayed as an image, as initially in the CSV file they were present as 784 columns of pixel data. So we convert it to 28×28 pixels.

```
word_dict =
{0:'A',1:'B',2:'C',3:'D',4:'E',5:'F',6:'G',7:'H',8:'I',9:'J',10:'K',11:'L',12:'M',13:'N',14:'O'
,15:'P',16:'Q',17:'R',18:'S',19:'T',20:'U',21:'V',22:'W',23:'X', 24:'Y',25:'Z'}
```

All the labels are present in the form of floating point values, that we convert to integer values, & so we create a dictionary word_dict to map the integer values with the characters.

Plotting the number of alphabets in the dataset

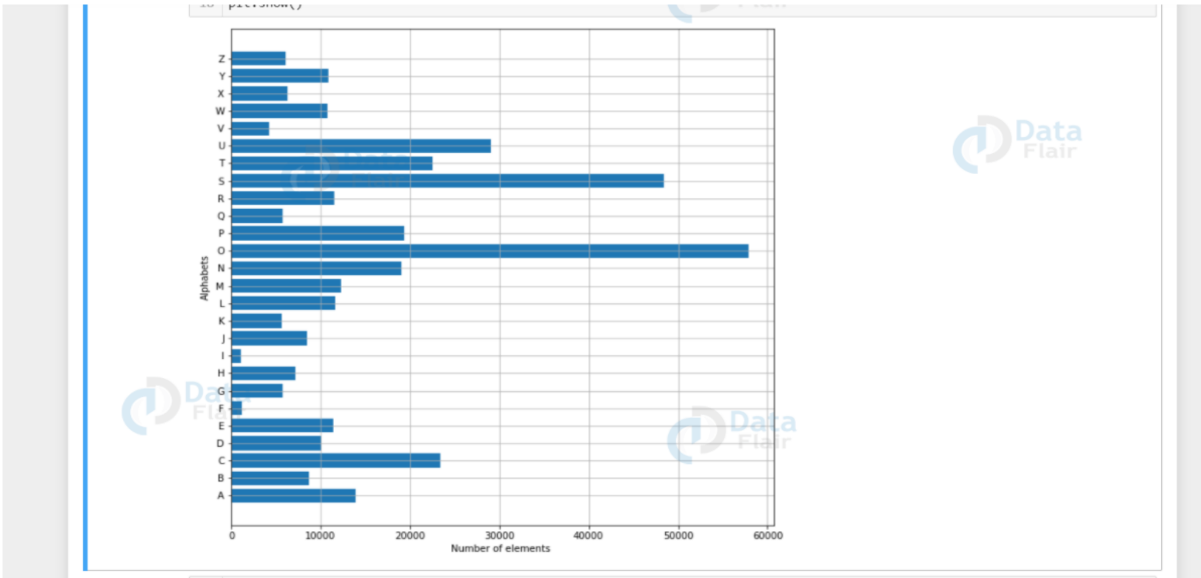
```
y_int = np.int0(y)
count = np.zeros(26, dtype='int')
for i in y_int:
    count[i] +=1

alphabets = []
for i in word_dict.values():
    alphabets.append(i)

fig, ax = plt.subplots(1,1, figsize=(10,10))
ax.barh(alphabets, count)

plt.xlabel("Number of elements ")
plt.ylabel("Alphabets")
plt.grid()
plt.show()
```

we are only describing the distribution of the alphabets.

Firstly we convert the labels into integer values and append into the count list according to the label. This count list has the number of images present in the dataset belonging to each alphabet.

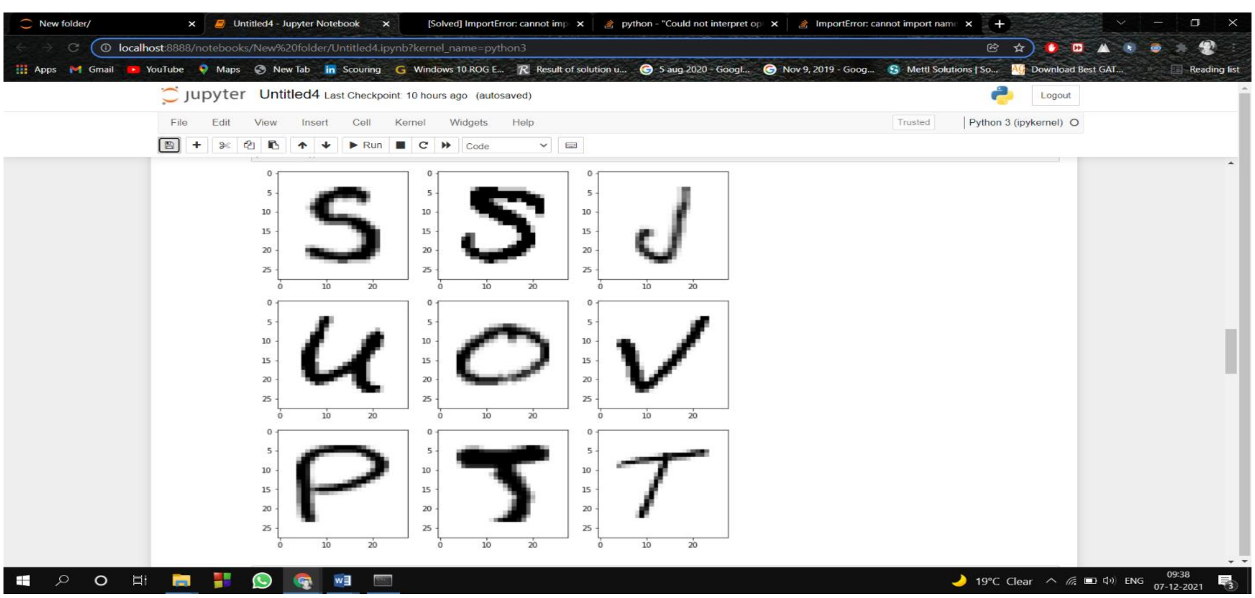
Now we create a list – alphabets containing all the characters using the values() function of the dictionary.

Now using the count & alphabets lists we draw the horizontal bar plot.



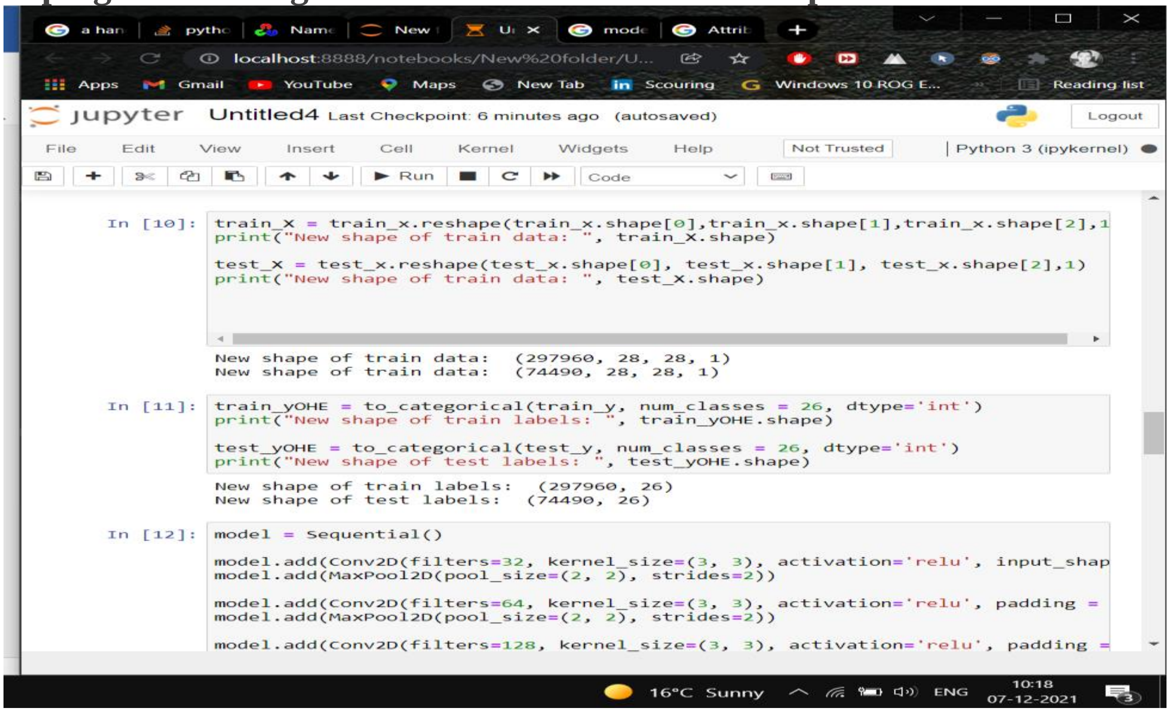
Shuffling the data

- we shuffle some of the images of the train set.
- The shuffling is done using the shuffle() function so that we can display some random images.
- We then create 9 plots in 3×3 shape & display the thresholded images of 9 alphabets.



Data Reshaping

Reshaping the training & test dataset so that it can be put in the model

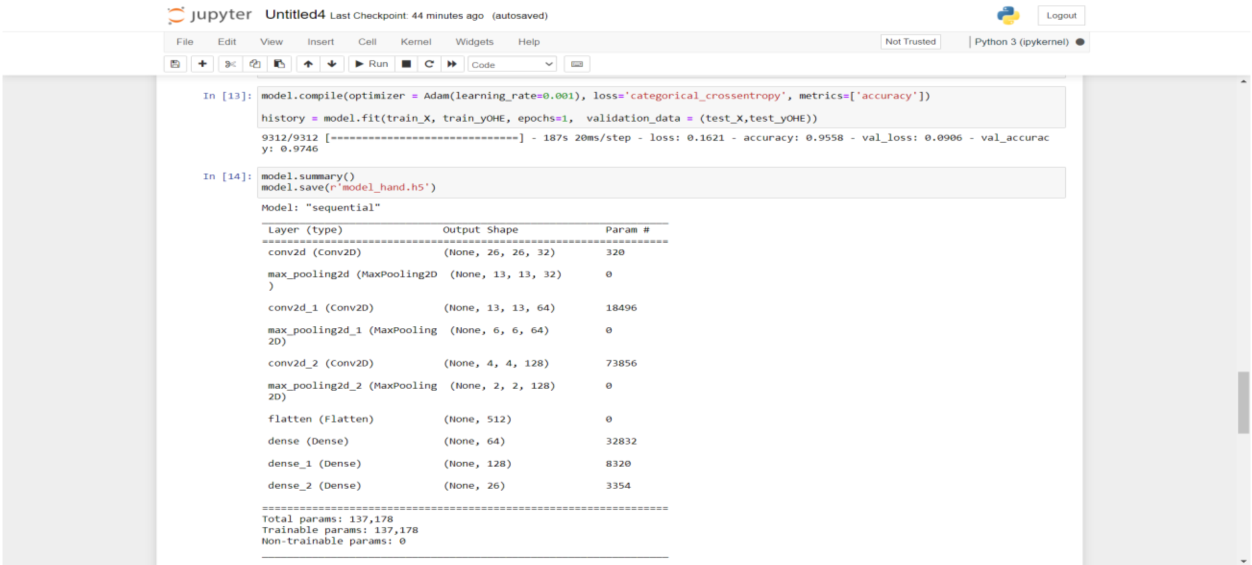


Feature Extraction As individual characters has been separated, character image can be re sized to 15 x 20 pixels. If the features are extracted accurately then the accuracy of recognition ismore.Here we have use the 15 x 20 means 300 pixels as it is for feature vector.

Then convert the single float values to categorical values. This is done as the NN model takes input of labels & generates the output as a vector of probabilities.

Compiling & Fitting Model

- we are compiling the model, where we define the optimizing function & the loss function to be used for fitting.
- The optimizing function used is Adam, that is a combination of RMSprop & Adagram optimizing algorithms.
- The dataset is very large so we are training for only a single epoch, however, as required we can even train it for multiple epochs (which is recommended for character recognition for better accuracy).

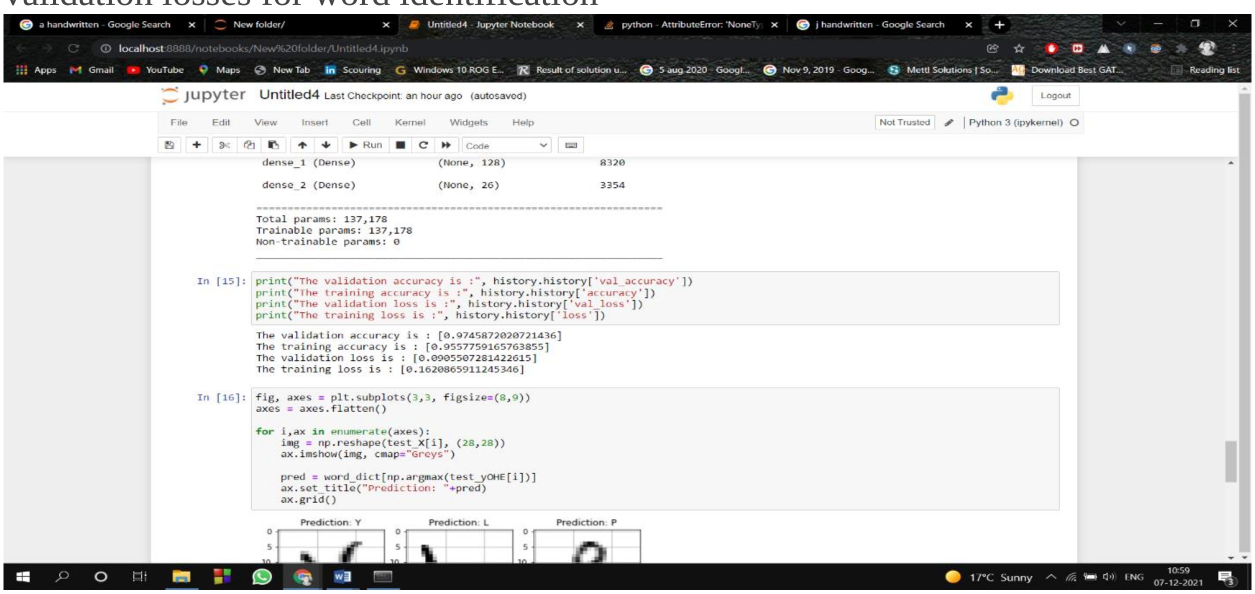


we are getting the model summary that tells us what were the different layers defined in the model & also we save the model using **model.save()** function.

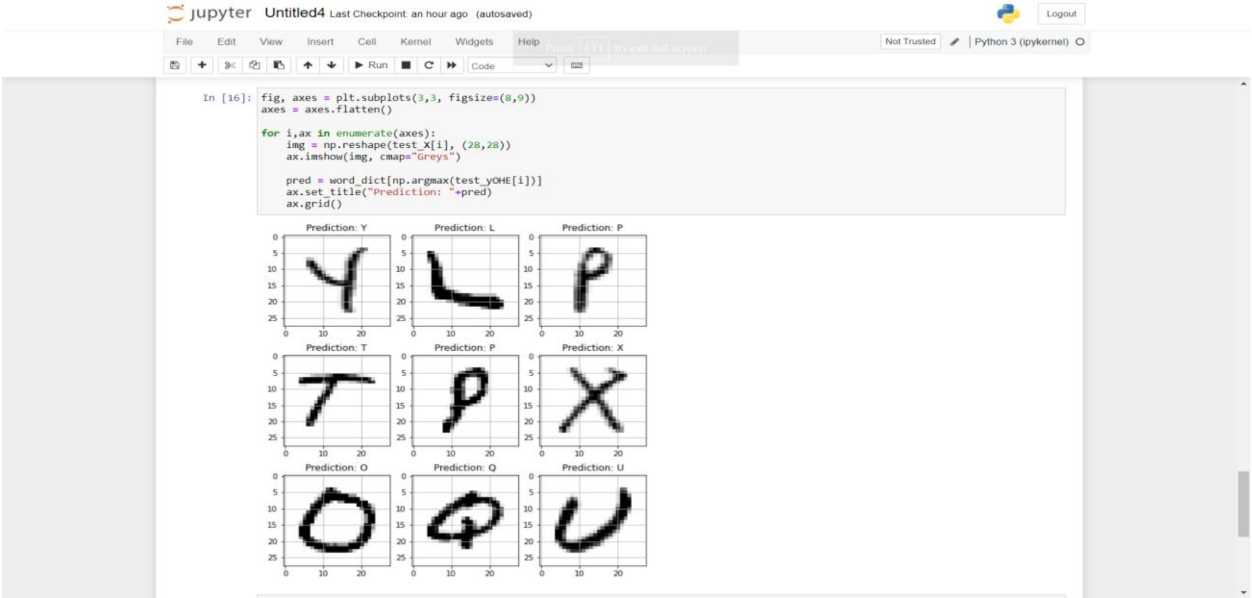
Getting the Train & Validation Accuracies & Losses

```
print("The validation accuracy is :", history.history['val_accuracy'])
print("The training accuracy is :", history.history['accuracy'])
print("The validation loss is :", history.history['val_loss'])
print("The training loss is :", history.history['loss'])
```

we print out the training & validation accuracies along with the training & validation losses for word identification



we are creating 9 subplots of (3,3) shape & visualize some of the test dataset alphabets along with their predictions, that are made using the **model.predict()** function for text recognition and changes after every run .



Doing Prediction on External Image

img = cv2.imread(r'D:\test\j.jpg')
img_copy = img.copy()
img = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
img = cv2.resize(img, (400,440))

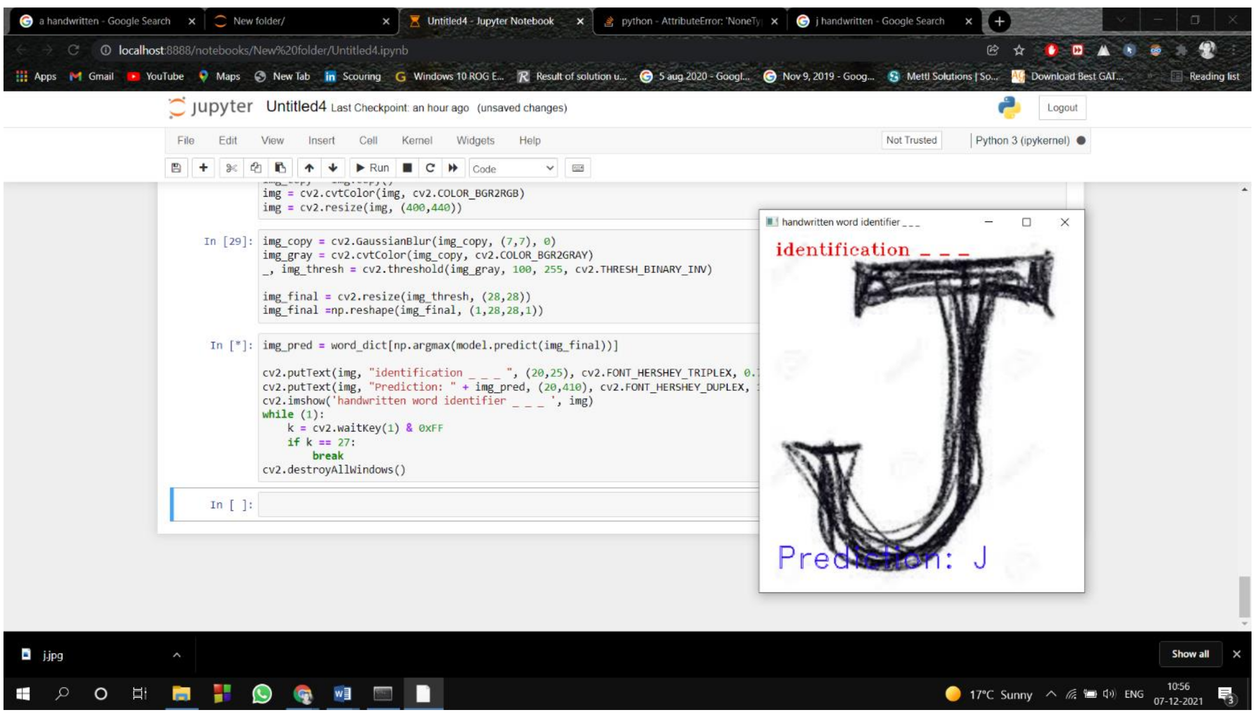
- we have read an external image that is originally an image of alphabet ‘J’ and made a copy of it that is to go through some processing to be fed to the model for the prediction that we will see in a while.
- The img read is then converted from BGR representation (as OpenCV reads the image in BGR format) to RGB for displaying the image, & is resized to our required dimensions that we want to display the image in.

- Now we do some processing on the copied image (img_copy).
- We convert the image from BGR to grayscale and apply thresholding to it. We don’t need to apply a threshold we could use the grayscale to predict, but we do it to keep the image smooth without any sort of hazy gray colors in the image that could lead to wrong predictions.

- The image is to be then resized using **cv2.resize()** function into the dimensions that the model takes as input, along with reshaping the image using **np.reshape()** so that it can be used as model input.
- Now we make a prediction using the processed image & use the np.argmax() function to get the index of the class with the highest predicted probability. Using this we get to know the exact character through the word_dict dictionary.
- This predicted character is then displayed on the frame.
- Here we are setting up a waitKey in a while loop that will be stuck in loop until Esc is pressed, & when it gets out of loop using cv2.destroyAllWindows() we destroy any active windows created to stop displaying the frame.

Result:-

“Handwritten word identifier ”is aimed at recognizing the handwritten characters. The “handwritten word identifier System ”is implemented using a neural network.in this system original image is converted into gray scale image then converted from BGR representation (as OpenCV reads the image in BGR format) to RGB for displaying the image, & is resized to our required dimensions that we want to display the image in a prediction using the processed image, successfully developed Handwritten word identifier with Python, Tensorflow, and Machine Learning libraries.Handwritten characters have been recognized with more than 97% test accuracy..



Conclusion and Future scope:-

Handwritten word identifier is one of the interesting fields of research in the image processing. Though lot of work has been done still it has got many opportunities to do the work in this field. Reasons for this are, in the most of the languages lack of availability of standard datasets and accuracy rate obtained. By using good combinations of feature extraction methods and classifiers it is possible to achieve the good results and the application of neural network and main of our project is to recognize the word and so there is kaggle dataset for the sample for training the model and our future work is to add mnist dataset in this same execution program to identify the digit data using neural network.

REFRENCES:-

1.https://www.researchgate.net/publication/315053242_Handwritten_Character_Recognition_in_English_A_Survey

2.https://www.ijert.org/research/offline-handwritten-word-recognition-a-survey IJERTCONV3IS19112.pdf

3.https://www.researchgate.net/publication/323547763_Handwritten_Character_Recognition_HCR_Using_Neural_Network

4.https://thesai.org/Downloads/Volume11No7/Paper_19-Handwriting_Recognition_using_Artificial_Intelligence.pdf

5. B. Gatos, I. Pratikakis, A.L. Kesidis, S.J. Perantonis, "Efficient OffLine Cursive Handwriting Word Recognition", Tenth International Workshop on Frontiers in Handwriting Recognition, Oct. 2006

6.] Ankush Acharyya, Sandip Rakshit, Ram Sarkar, Subhadip Basu, Mita Nasipuri, "Handwritten Word Recognition Using MLP based Classifier: A Holistic Approach", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 2, March 2013

7. Rodolfo Luna-Perez, Pilar Gomez-Gil, "Unconstrained Handwritten Word Recognition Using a Combination of Neural Networks", ISBN: 978-988-17012-0-6, WCECS 2010

8.Shaolei Feng Nicholas R. Howe R. Manmatha, "A Hidden Markov Model for Alphabet-Soup Word Recognition", Dept. of Computer Science, University of Massachusetts, Amherst, 2008

9. Alex Graves, Jurgen Schmidhuber, "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks", In NIPS, PP. 545-552,2008 10.

M.Sivasankari , Dr.S.P.Victor" Handwritten Text Recognition"ISSN 0975-9646
11. Annadurai, S. and Shanmugalakshmi, R. "Fundamentals of digital image processing", Pearson Education, 2007.
12. Megha Agarwal, Shalika, Vinam Tomar, Priyanka Gupta" Handwritten Character Recognition using Neural Network and Tensor Flow" ISSN: 2278-3075
13. Sara Aqab , Muhammad Usman Tariq" Handwriting Recognition using Artificial Intelligence Neural Network and Image Processing"
14. Monica Patel , Shital P. Thakkar" Handwritten Character Recognition in English: A Survey" ISSN (Online) 2278-1021
15. M S Patel ,Rohith Kumar" Offline Handwritten Word Recognition: A Survey" ISSN: 2278-0181