Machine Learning Approach for Translating Handwritten Document to Digital Form

Submitted By

Tuhin Saha(16900118029)

Soumabha Majumdar(16900118052)

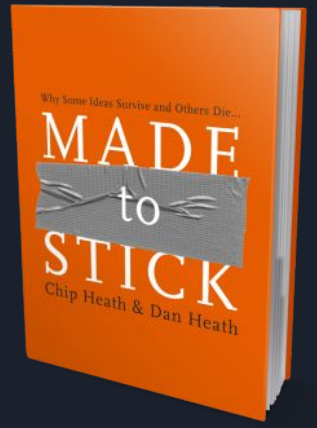
Shreya Rai(16900118059)

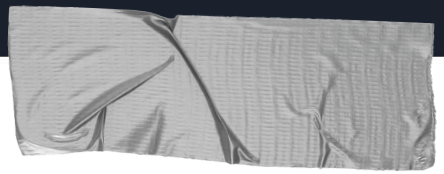
Sweta Mishra(16900118038)

Siddhant Sharma(16900118055)

Department- CSE, Semester- 7th Subject: Project-II (PROJ- CS781)

Under the guidance of Prof. Somen Hati





1. Intro

Handwritten papers are not perfect, and for one, they are difficult to read through. Handwriting recognition has been one of the challenging research areas in the field of image processing and pattern recognition in recent years. It contributes immensely to the advancement of the 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 pre-processing time while providing higher recognition accuracy.

Purpose Of Study

Handwriting recognition is a challenging task because of many reasons. The primary reason is that different people have different styles of writing. The secondary reason is there are a lot of characters like capital letters, small letters, digits and special symbols. Thus, a large dataset is required to train the system. Optical character recognition (OCR) is usually referred to as an off-line character recognition process to mean that the system scans and recognizes static images of the characters.

Fig 1

Literature Overview

Exhaustive research is being done in the field of image processing and Machine Learning towards handwritten recognition. We have observed that Scale Conjugate algorithm has achieved an accuracy of 95% in classifying the characters and turned out to be better than Resilient Back propagation algorithm. The proposed ANN model accepts the handwritten text as input, processes it, extracts the relevant optimal features and applying one of the algorithms recognizes the characters. Multiple characters input in a single image, tilt image and rotated image is also being tried applying additive image processing algorithms. We have used Intelligent Word Recognition to identify whole word in a document. Segmentation algorithms are being used to separate cursive and joined handwriting. Binary Segmentation Algorithm has given acceptable performance in extracting individual characters from words. We have designed a Machine Learning model for recognizing handwritten characters on form document.

- Line segmentation: Horizontal projection of a document image is most commonly used to extract the lines from. Only lines are extracted or differentiated from the document.
- Word segmentation: A process of dividing a string into its component words. It is a process of parsing concatenated text to infer where word breaks exist.
- Character segmentation: A process where only characters are extracted from word. It is a difficult step of OCR systems which decomposes the images into classifiable units called character.

An Offline handwriting recognition is being designed which is the automatic transcription by computer of handwriting, where only the image of the handwriting is available. Offline handwriting is thus distinguished from online handwriting, where the path of the pen is measured by a device such as a digitizing tablet. A host of applications of off- line handwriting can be envisaged, including document transcription, automatic mail routing, and machine pro-cessing of forms, checks, and faxes.

Problem Definition and Overview

- It's easier than ever to jot down digital notes on computers and phones, but many people still prefer the traditional feeling of writing with ink It's easier than ever to jot down digital notes on computers and phones, but many people still prefer the traditional feeling of writing with ink on paper. (After all, this method served us well for hundreds of years of human history.) The problem is that you can't organize and search through handwritten notes the way you can with files on a digital device. Or can you? A number of smart devices offer to digitize your scribblings, either as you write or shortly afterward. In addition to storing images of your notebook pages in electronic form, some of these hardware and software packages actually convert your writings into searchable text. Our project utilises various software and machine learning paradigms to convert such handwritten notes into digital searchable form. Those paradigms will be introduced and illustrated in following chapters.on paper. (After all, this method served us well for hundreds of years of human history.) The problem is that you can't organize and search through handwritten notes the way you can with files on a digital device. Or can you? A number of smart devices offer to digitize your scribblings, either as you write or shortly afterward. In addition to storing images of your notebook pages in electronic form, some of these hardware and software packages actually convert your writings into searchable text. Our project utilises various software and machine learning paradigms to convert such handwritten notes into digital searchable form. Those paradigms will be introduced and illustrated in following chapters.

Feasibility Study

Machine Learning is a data analytics technique that teaches computer to do what comes naturally to humans and animals that learn from experience. Machine Learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Machine Learning has become a key technique for solving problems in areas, such as

- **Computational finance: credit scoring and algorithmic trading**
- **Image processing and computer vision: face recognition, motion detection, and object detection**
- **Computational biology: tumour detection, drug discovery, and DNA sequencing**
- **Energy production: price and load forecasting**
- **Automotive, aerospace and manufacturing: predictive maintenance**
- **Natural language processing: voice recognition applications.**

MATLAB lets you: • Extract features from signals and images using established manual and automated methods • Compare approaches such as logistic regression, classification trees, support vector machines, ensemble methods, and deep learning. • Apply AutoML and other model refinement and reduction techniques to create optimized models. • Integrate Machine Learning models into enterprise systems, clusters, and clouds, and target models to real-time embedded hardware. • Perform automatic code generation for embedded sensor analytics. • Support integrated workflows from data analytics to deployment. MathWorks released 2018a with a range of new capabilities in MATLAB and Simulink. R2018a includes two new products. Predictive Maintenance

A. Computer Vision System Toolbox (Version 10.2)

Computer Vision System Toolbox provides algorithms, functions, and apps for designing and simulating computer vision and video processing systems. Functions like feature detection, extraction, and matching, as well as object detection and tracking are adopted for detecting characters in handwritten text. For 3-D computer vision, the system toolbox supports single, stereo, and fish eye camera calibration; stereo vision; 3-D reconstruction; and 3-D point cloud processing.It can train a custom detector using ground truth labelling with training frameworks such as Faster R-CNN and ACF. In addition, it can classify image categories and perform semantic segmentation. Algorithms are available as MATLAB functions, System objects, and Simulink blocks. For rapid prototyping and embedded system design, the system toolbox supports fixedpoint arithmetic and C-code generation.

B. Image Processing Toolbox (Version 10.2)

Image Processing Toolbox provides a comprehensive set of reference-standard algorithms and workflow apps for image processing, analysis, visualization, and algorithm development. Can perform image segmentation, image enhancement, noise reduction, geometric transformations, image registration, and 3D image processing. It lets to automate common image processing workflows. In addition, it can interactively segment image data, compare image registration techniques, and batch-process large datasets. Visualization functions and apps

C. Neural Networks Toolbox (Version 11.1)

Neural Network Toolbox provides algorithms, pre-trained models, and apps to create, train, visualize, and simulate both shallow and deep neural networks. It can perform classification, regression, clustering, dimensionality reduction, time-series forecasting, and dynamic system modelling and control. Deep learning networks include convolutions neural networks (ConvNets, CNNs), directed acyclic graph (DAG) network topologies, and auto encoders for image classification, regression, and feature learning. For time-series classification and regression, the toolbox provides long short-term memory (LSTM) deep learning networks. Intermediate layers and activations, modify network architecture, and monitor training progress can be visualized. For small training sets, deep learning can be applied by performing transfer learning with pre-trained deep network models (including Inception-v3, ResNet-50, ResNet-101, GoogleNet, AlexNet, VGG-16, and VGG-19) and models imported from TensorFlow R - Keras or Caffe. To speed up training on large datasets, computations and data can be distributed across multicore processors and GPUs on the desktop (with Parallel Computing Toolbox), or scale up to clusters and clouds, including Amazon EC2 P2, P3, and G3 GPU instances (with MATLAB R Distributed Computing Server). The proposed work adopts the above-mentioned toolboxes to design neural network model. The designed model has g

System Analysis

MATLAB provides a platform for the creation of the Graphical user interface through “Guide”. Guide stands for Graphical User Interface Development Environment. It contains all the required tools for the creation. It provides the tools to design user interfaces and create custom apps. It also provides pointand-click control of a software applications, eliminating the need for others to learn a language or type commands in order to run the application. Few tools are used for the creation of the front which is depicted in fig. 5. 1) Capital letter detection: The input text image is taken. The image is converted to a grayscale. Line is split followed by letter segmentation. 2) Small letter detection: A template is created at the beginning

Fig. 3. Neural Network Toolbox

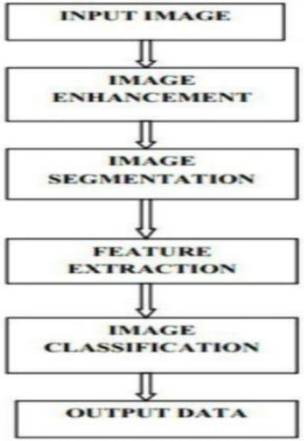


Fig. 4. Flow of the process

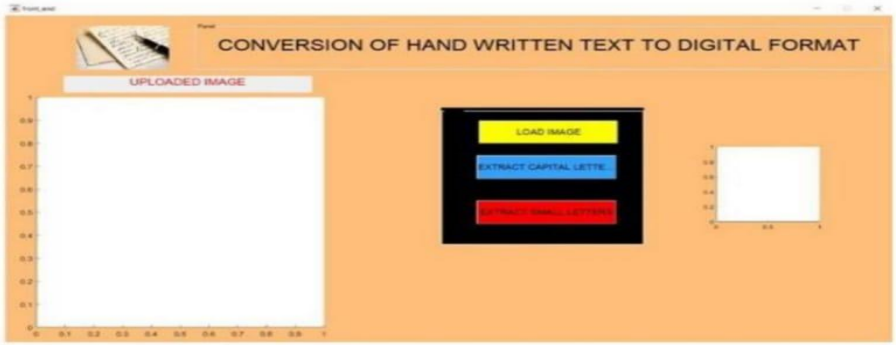


Fig. 5. Front end of the designed tool

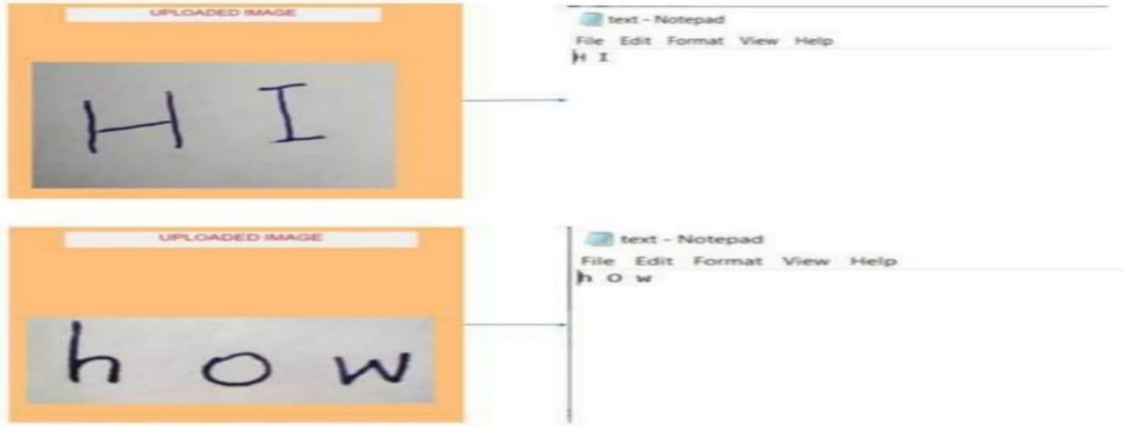
A. Dataset Images are obtained from MNIST dataset for experimental analysis. The MNIST database is a large database of handwritten digits that is commonly used for training various image processing systems. The database is also widely used for training and testing in the field of Machine Learning. It was created by remixing the samples from NIST's original datasets. The MNIST database contains 60,000 training images and 10,000 testing images. Half of the training set and half of the test set were taken from NIST's training dataset, while the other half of the training set and other half of the test set were taken from NIST's testing dataset. The dataset is available in the MATLAB Matrix format. It is an array of 124000x784 with each row representing one image. The original size of the image is 28x28 pixels which is converted into a 1x784 array and stored.

B. Creation of Neural Network Classification and identification of images are processed using Neural Network. Neural network is a branch of Artificial Intelligence that imitates the biological processing function of the brain. Neural network has been implemented in various applications and one of the applications is handwritten recognition system. Handwritten is the art of an individual which is controlled by the function of the brain. Every individual has his or her own style of writing. Hence, reading the handwriting is sometimes quite difficult. Many researches have been done in this area and yet still continuing. The potential of NN attracted many researchers to develop and integrate NN in their applications and one such area of interest is handwritten recognition. Handwriting is a series of complex actions that involves human nerve system, physical, emotion and natural behaviour. The innovation of input devices such as digitizer, enable the computer to capture the handwriting while it is being applied on a paper. This approach is called online method. It can also be converted into a digital form using scanner or any OCR devices. For the creation of neural network, the neural network fitting tool has been used with Sigmoid hidden neurons. Two networks are being created and experimented. Network1 was designed using 10 neurons and Network2 with 15 neurons.

C. Error Histogram Error histogram is the histogram of the errors between target values and predicted values after training a Feed Forward Network. These error values indicate how predicted values are differing from the target values; hence these can be negative. The total error range is divided into 30 smaller bins here. Yaxis represents the number of samples from the dataset, which lies in a particular bin. For example, at the mid of the plot, a bin is corresponding to the error of 0.001502 and the height of that bin for training dataset lies below but near to 150 and validation and test dataset lies between 150 and 200.

D. Optical Character Recognition (OCR) During experimental analysis, along with MNIST data set as input images, input is also generated using OCR and tested. An OCR system is a computerized scanning system enabling to scan text documents into an electronic computer file which can then be edited using a word processor on the computer. OCR is the machine recognition of printed text character OCR software works with scanner to convert printed characters into digital text, allowing to search for or edit document in a word processing program. Widely used form of data entry from printed paper data records- passport documents computerized receipts, invoices, bank statements, business cards, mail, printouts of static data, or any

Optical character reader



Output of proposed work

Software and Hardware Requirement

Hardware Interfaces : For the product to run, you shall need an Windows/Mac device. To get accurate results, your device should have a high resolution camera. Also, since image processing is resource-hungry, your device should have modern RAM and CPU. Since neither the application nor the web portal have any designated hardware, it does not have any direct hardware interfaces. The physical camera is managed by the camera interface in the smart phone and the hardware connection to the Internet is managed by the underlying operating system on the computer and the web server. Software Interfaces As it has already been stated, product shall run on only Windows/Mac devices, so it shall not need any specific software from the user's side other than it. The application communicates with the web server in order to get lottery results. However, product shall depend on some libraries and tools to be developed. User interface module to be implemented shall interact with user and according to its input, shall activate image processing module. Image processing module shall be implemented with MATLAB
Web module shall connect to Milli Piyango Web Site and shall
gather related data.

BIBLIOGRAPHY

Books : • OCR with OpenCV, Tesseract, and Python OCR Practitioner Bundle by Adrian Rosebrock, PhD Abhishek Thanki Sayak Paul Jon Haase. • Machine Learning Mastery With Python by Jason Brownlee.

Websites : • https://towardsdatascience.com/ • https://machinelearningmastery.com/ • https://www.w3schools.com/python/ • https://www.tutorialspoint.com/python/index.htm • https://www.geeksforgeeks.org/opencv-python-tutorial/

A Project Presentation To be submitted in the partial fulfillment of the requirements

For the degree of

Bachelor of Technology in Computer Science and Engineering

Department of Computer Science and Engineering,

Academy of Technology

THANK YOU