Machine Learning Approach for Translating Handwritten Document to Digital Form

[2] Sowmya Hegde, [3]Shreyashree A V,
[4]Bimba Prasad, [5]Chinmayi,
Final Year Students, Department of Computer Science ,
Malnad College of Engineering, Hassan,
Karnataka, India

[1] Sunitha P
Assistant Professor,
Department of Computer Science and Engineering,
Malnad College of Engineering, Hassan,
Karnataka, India

Abstract— Computers and phones may be more ubiquitous than ever, 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. Despite the availability of various technological writing tools, many people still choose to take their notes traditionally: with pen and paper. However, there are certain pitfalls in traditional way of handwritten text. It is difficult to store and access physical documents in an organised manner, search through them efficiently and to share them with others. Thus, handwriting recognition is the ability to interpret intelligible handwritten input from sources such as paper documents, touch-screens and other devices into digital form. A handwriting recognition system handles formatting, performs correct segmentation into characters, and finds the most plausible words. Hence, translating the handwritten characters to the digital format is gaining more popularity. With time the text on the paper will fade away but a file stored on a computer will be lost only if it is deleted. Storing any handwritten document in a digital format has gained prime importance.

Once the handwritten document is given as the input in the form of a high definition image, it segments each character in the image and identifies the letters. Further, the letters are identified and then goes on to detect the words in the image. This is performed with the aid of Machine Learning algorithms based on the training it has got from the training data. The expected output is to get a word document format of the given input image. The system can be trained by large data set of images that show the various styles and shapes in which people write. Machine Learning plays a very important role in training the system with huge data. This can be further used in organizations and companies that store important documents only in written format. It becomes easier and faster to complete the work with such a system available at hand.

Keywords - Machine Learning, Image Processing, Feature Extraction, Neural Network, Computer Vision, Segmentation

I. INTRODUCTION

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[1] in recent years. It contributes immensely to the advancement of the automation process and improves the interface between man and machine in numerous applications[2]. Several research works have been focusing on new techniques and methods that would reduce preprocessing time while providing higher recognition accuracy[3,4]. 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.

Handwriting data is converted to digital form either by scanning the writing on paper or by writing with a special pen on an electronic surface. The two approaches are distinguished as offline and on-line handwriting, respectively. In the on-line case, the two-dimensional coordinates of successive points of the writing as a function of time are stored in order. In the off-line case, only the completed writing is available as an image. Figure 1 shows the analysis of the two cases. The recognition rates reported are much higher for the on-line case in comparison with the off-line case. Off-line systems are less accurate than online systems. However, they are now good enough that they have a significant economic impact for specialized domains such as interpreting hand-written postal addresses on envelopes and reading courtesy amounts on bank checks. The success of on-line systems makes it attractive to consider developing, off-line systems that first estimate the trajectory of the writing from offline data and then use on-line algorithms.

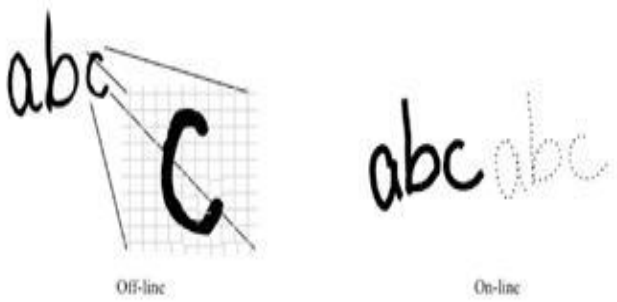

Fig. 1. Analysis of two cases

II. LITERATURE SURVEY

Exhaustive research is being done in the field of image pro-cessing and Machine Learning towards handwritten recognition.

[12]Authors have proposed a 3 layer ANN with two different learn-ing algorithms. They 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. Kavitha ans Shamini 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 Al-gorithm has given acceptable performance in extracting individual characters from words[17].

[3] Authors have designed a Machine Learning model for rec-ognizing handwritten characters on form document. The learning

model is based on Convolution Neural Network (CNN) as powerful feature extraction model and Support Vector Machines (SVM) as a high-end classifier. The proposed method is found to be more efficient than modifying CNN with complex architecture. The proposed system has achieved a recognition rate of 98.85% on numeral characters, 93.05% on uppercase characters, 86.21% on lowercase characters, and 91.37% on the merger of numeral and uppercase characters. The pre-processing, segmentation and character recognition are integrated into one system. The output of the system is converted into an editable text. The system gives an accuracy rate of 83.37% on ten different test form document.

The handwritten database adopted for experiment contains im-ages of 26 training images, 26 validation images, and 26 test im-ages. The offline handwriting recognition is the technique involves image capture, enhancement and recognition. The image capture step involves the scanning of the image with 300 dots per inch resolution and 256 gray level scaling.

Similarly, [19]Authors have proposed a light weight structure using CNN for mobile devices. Mobiles are used to collect data, process it and generate data for training and testing the CNN.They were able to generate 400 variety of images and among 50000 images 75% data is selected for training and 25% for testing. Their proposed model was able to achieve effective classification performance of 93.3% accuracy. The model was tested with Alexnet and Googlenet data also and Googlenet gave hihest accuracy than Alexnet. [18] Authors have applies decision trees along with Neural Network to classify characters and words. Converted text is further transformed to voice format achieving 92.7% accuracy.

In [16], authors have proposed an algorithm using Radon Trans-form and Back Propagation Network(RTBPN). The gray scale image is scanned and the resultant gray scale image is normalized. Thresholding techniques are used in this normalization process. Here, the first step of segmentation is line segmentation and word segmentation followed by baseline estimation [6-8]. The three baselines are the base parameters for the future analysis of the separated word. This baseline estimation is done after splitting the entire image into sub-images in the following sequence. Radon transform is the Handwritten Text to Digital Text Conversion Using RTBPN. 499 fundamental tools are being used in the approach. The Radon transform is used to detect linear trends in images. The principal direction of the image is first estimated from the Radon transform of a disk shape area of the image . The cropped word is then further cropped into ‘n’ reasonable sub words. For each sub-word, the radon transform is applied separately to calculate the shear angle.

[15]Authors have made a detailed analysis of various tech-niques for translating Handwritten Off-Line Cursive Words. It was observed that holistic approach is more suitable for applications where three segmentation based approaches for cursive handwriting recognition was proposed by the the lexicon is statically defined. Explicit segmentation based approach was computationally complex than implicit segmentation but gives slightly better results than less complex implicit based Segmentation approach. Segmentation process included the following processes:

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.

[6,13]Authors have made a novel attempt in recognizing hand-written characters by using multi-layer Feed Forward NN. During the process of recognizing characters, feature extraction phase is exempted and has achieved a better performance than extracting feature based approach. Of the several neural networks architec-tures, the propose model with input nodes of 10 and two layer NN , each layer having 100 neurons , an accuracy of 90% accuracy is achieved in identifying handwritten text.

[8]An Off-line handwriting recognition is being designed which is the automatic transcription by computer of handwriting, where only the image of the handwriting is available. Off-line handwriting is thus distinguished from on-line 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. Other systems to recognize off-line handwriting have been produced, but most are limited to digit recognition or small vocabulary transcription problems, such as the postal or check-reading applications where the context or additional knowledge, e.g., the zip code, limits the vocabulary considerably. The system has been designed to tackle the large-vocabulary task of text transcription. The work has been carried out on a publicly-available database of writing from a single author. As such the envisaged application is for the transcription of documents by one writer—either for personal notes, off-line data entry, or potentially for historical document transcription. All of the techniques used are applicable to a writer independent transcription system that could be used to transcribe incoming mail, or faxes, address blocks, or checks.

The next stage in the process of deducing word identities from handwriting is to recognize what is represented by the frames of data created in the previous section. A variety of pattern recognition methods is available, and many have been used for hand for handwriting recognition by other authors. There are several established methods of estimating a sequence of probabilities from a sequence of data which have been applied in the fields of both speech and handwriting recognition. From the literature, two main methods emerge. Discrete or continuous hidden Markov models, and neural network/HMM hybrids [15], have been successfully applied in speech, on-line handwriting, and off-line handwriting recognition. For off-line recognition, where time information is not available, the x-axis is generally divided up and processed left-to-right over time. Among the neural-network approaches, feed forward and recurrent network approaches can be distinguished. The latter have been successful in speech recognition, but have not previously been applied in handwriting recognition.

[11]Prathibha et. al have developed a Neural Network(NN) using LabView which classifies an recognises characters an symbols. An accuracy of 70-80% has been achieved. NN has helped to decide the threshold value in classification process. Input data of the text document is accepted through OCR. Similarly, [12] authors have experimented by applying Back Propagation algorithm in identifying handwritten text for authentication. OCR is used to scan the input image in off-line mode and digital pen in on-line mode. The scanned image is processed, vital features are extracted and converted image to matrix form is given as input to NN. Experiment is conducted with NN starting from 1 layer. With more number of layers, hidden units and minimizing error threshold, more accurate authentication is being achieved. [14] Authors have integrated feature extraction and classification phases to enhance the character recognition accuracy. Classifiers like template matching, Feed Forward NN(FFNN), Nearest Neighbour NN, Radial basis Function NN are being models for experiment and FFNN is found to obtain the highest accuracy of 94

III. IMPLEMENTATION

Machine Learning is a data analytics technique that teaches com-puters 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: tumor detection, drug discovery, and DNA sequencing

Energy production: price and load forecasting

Automotive, aerospace, and manufacturing: predictive main-tenance

Natural language processing: voice recognition applications

Machine Learning algorithms find natural patterns in data that generate insight and help to make better decisions and predictions. They are used every day to make critical decisions in medical diagnosis, stock trading, energy load forecasting, image processing and more.

Thus, it is needed to harness the power of Machine Learning to use data to make better decisions. MATLAB makes Machine Learning easy with tools and functions for handling big data, as well as apps to make Machine Learning accessible. MATLAB is an ideal environment for applying Machine Learning to data analytics.

With MATLAB, engineers and data scientists have immediate access to pre-built functions, extensive toolboxes, and specialized apps for classification, regression, and clustering.

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 Toolbox for designing and testing condition monitoring and predictive maintenance algorithms, and Vehicle Dynamics Blockset for modeling and simulating vehicle dynamics in a virtual 3D environment are the new features of new version of MatLab.

The MatLab toolbox provides supervised and unsupervised Ma-chine Learning algorithms, including Support Vector Machines (SVMs), boosted and bagged decision trees, k-nearest neighbor, k-means, k-medoids, hierarchical clustering, Gaussian mixture mod-els, and hidden Markov models. Many of the statistics and Machine Learning algorithms can be used for computations on data sets that are too big to be stored in memory. Few of the toolboxes used in the proposed work for experimental purpose are mentioned below.

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.

Algorithms for deep learning and Machine Learning enables to detect faces, pedestrians, and other common objects using pre-trained detectors. It can train a custom detector using ground truth labeling 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 fixed-point 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 pro-cessing, analysis, visualization, and algorithm development. Can perform image segmentation, image enhancement, noise reduction, geometric transformations, image registration, and 3D image pro-cessing. 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. Visualiza-tion functions and apps helps to explore images, 3D volumes, and videos; adjust contrast; create histograms; and manipulate Regions of Interest (ROIs). It can accelerate algorithms by running them on multicore processors and GPUs.

Many toolbox functions support C/C++ code generation for desktop prototyping and embedded vision system deployment.

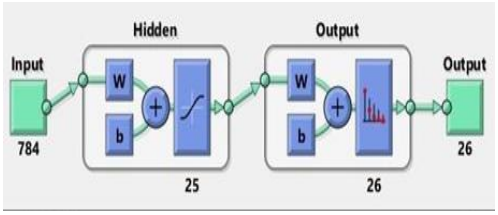

Fig. 2. Image Processing Toolbox

C. Neural Networks Toolbox (Version 11.1)

Neural Network Toolbox provides algorithms, pre-trained mod-els, 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 modeling and control.

Deep learning networks include convolutions neural networks (ConvNets, CNNs), directed acyclic graph (DAG) network topolo-gies, 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 perform-ing transfer learning with pre-trained deep network models (includ-ing 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 given satisfactory results in transforming handwritten text to digital form. Fig. 4 depicts the flow chart of the process involved in the proposed work.

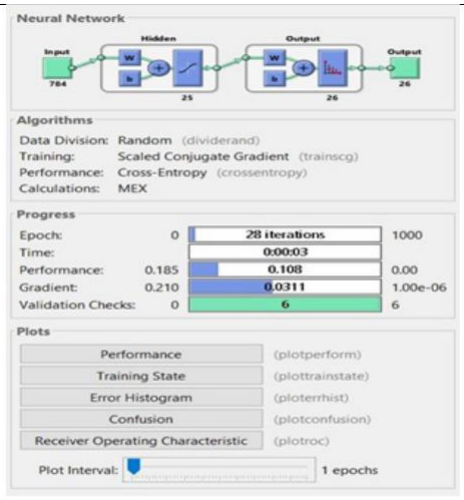

Fig. 3. Neural Network Toolbox

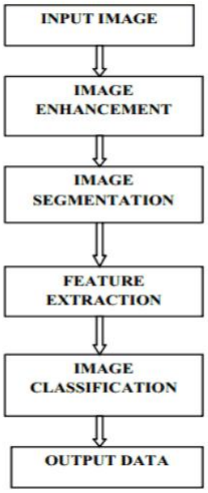

Fig. 4. Flow of the process

1) Capital letter detection: The input text image is taken as

tion. The image is converted to a grayscale

Line is split followed by letter segmentation. Once

of

he previously trained

2) Small letter detection: A template is created at the beginning

pped in

in the template and the value extracted from feature correlation is applied to identify the alphabet. After detection the letter is written into the output text file.

IV. Experiment and Result 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 point-and-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.

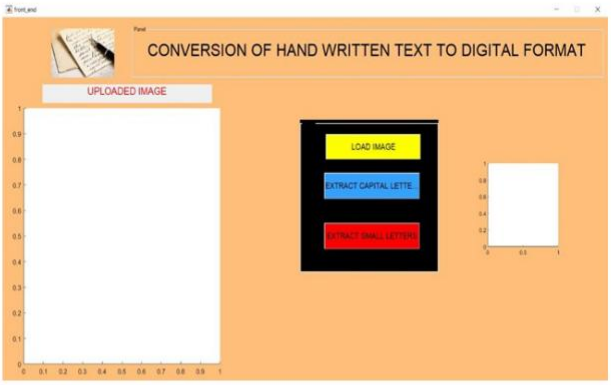

Fig. 5. Front end of the designed tool

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 re-mixing 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 In-telligence 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 behavior. 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.

In fitting tool, a set of numeric input is mapped to a set of numeric targets.

Training-70
Validation-15
Testing-15

By training a network, it grasps all the features of the samples. Network1 took 6 hours to train and Network2 took 27 hours with 73 iterations. Network2 gave acceptable performance in identifying handwritten text.

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. Y-axis 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. It means that many samples from different datasets have an error that lies in the following range. Zero error line correspond to the zero error value on the error axis. In this case zero, error point falls under the bin with center 0.001504.

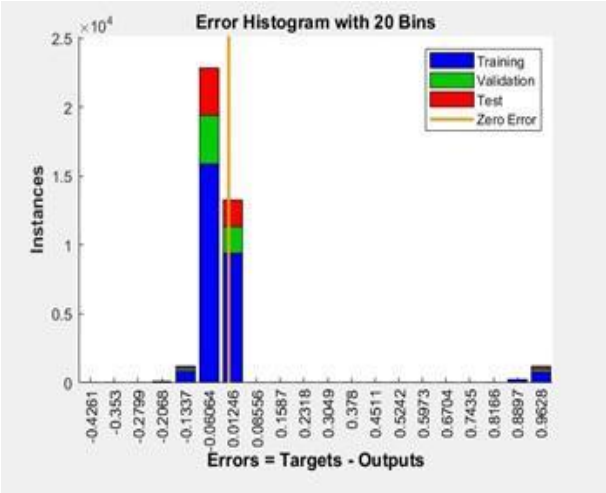

Fig. 6. Error Analysis

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 suitable documentation- it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, display online, and used in machine processes such as cognitive computing, machine translation, text- to- speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Fig. 7. Optical Character Reader

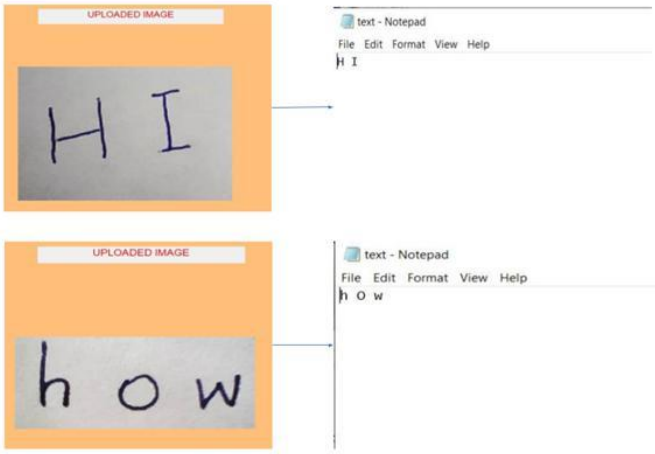

Fig. 8. Output of the Proposed work

V. CONCLUSION AND FUTURE WORK

The proposed handwritten recognition system recognizes the let-ters with acceptable accuracy. Extraction of small and capital letters is done separately. In future the small letters and capital letters can be integrated. There is a Scope for Recognition of digits and Special symbols. The information of the document can be edited more conveniently and can reuse the edited information as and when required. The grid infrastructure used in the implementation of Optical Character Recognition system can be efficiently used to speed up the translation of image based documents into structure documents that are currently easy to discover, search and process.

There is no out of box software for this kind of documents which can read handwriting automatically without human validation and training.

Font Independent OCR system could be developed by consid-ering the multiple font style in use. Our approach is very much useful for the font independent case. Because, for font or character size, it finds the string and the strings are parsed to recognize the character. Once character is identified, the corresponding character could be ejected through an efficient editor. Efforts have been taken

to develop a compatible editor for Tamil and English. There is heavy demand for an OCR system which recognizes cursive scripts and manuscripts like Palm Leaves. This actually avoids keyboard typing and font encoding too.

The most required application today is Speech recognition. The recognized Printed or Handwritten character could be recorded and through a voice synthesizer speech output could be generated. This would help the blind to send and receive information.

VI. REFERENCES

REFERENCES

[1] Joseph James S ,C.Lakshmi, UdayKiran P, Parthiban, An Efficient Offline Hand Written Character Recognition using CNN and Xgboost, International Journal of Innovative Technology and Exploring Engi-neering (IJITEE), April 2019

[2] Martin Rajnoha, Radim Burget, Handwriting Comenia Script Recog-nition with Convolutional Neural Network, ISBN: 978-1-5090-4911-0 (c) 2017 IEEE

[3] Darmatasia and Mohamad Ivan Fanany, Handwriting Recognition on Form Document Using Convolutional Neural Network and Support Vector Machines (CNN-SVM), Fifth International Conference on Information and Communication Technology (ICoICT), ISBN: 978-1-5090-4911-0 (c) 2017 IEEE

[4] Shubham Sanjay Mor, Shivam Solanki, Saransh Gupta, Sayam Dhin-gra, Monika Jain, Rahul Saxena, Handwritten Text Recognition : with Deep Learning and Android, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019 IEEE

[5] S. Mori, C.Y. Suen and K. Kamamoto, Historical review of OCR research and development, Proc. of IEEE, vol. 80, pp. 1029-1058, July 1992.

[6] J.Pradeep, E.Srinivasan, S.Himavathi, Neural Network based Hand-written Character Recognition system without feature extraction, In-ternational Conference on Computer, Communication and Electrical Technology – ICCCET 2011, 18th and 19th March, 2011

[7] N. Arica and F. Yarman-Vural, An Overview of Character Recognition Focused on Off-line Handwriting, IEEE Transactions on Systems, Man,And Cybernetics, Part C: Applications and Reviews, Vol.31 (2), pp. 216- 233. 2001.

[8] V.K. Govindan and A.P. Shivaprasad, Character Recognition – A review, Pattern Recognition, Vol. 23, no. 7, pp. 671- 683, 1990.

[9] J. U. Duncombe, "Infrared navigation Part I: An assessment of feasibility (Periodical style)", IEEE Trans. Electron Devices, vol. ED-11, pp. 34–39, Jan. 1959.

[10] Pratibha A. Desai, Sumangala N, Bhavikatti, Rajashekar Patil, Design an Simulation of Handwritten Text Recognition System, IJCET, 2013.

[11] Ahmed Mahi Obaid, Hazem M EI Bakry, M. A. Eldosuky, A.I. Shehab, Handwritten Text Recognition Based on Neural Network, IJARCST, Vol. 4, Issue 1, 2016.

[12] A. N. S. Chakravarthy, Penmesta V Krishna raja, P. S Avahani, Hand-written Text Image Authentication using Back Propagation, IJNSA, Vol. 3, No. 5, 2011

[13] Savitha Attigeri, Neural Network based Handwritten Character Recog-nition System, IJECS, Vol. 7, Issue 3, P.No. 23761-23768, 2015.

[14] J. Pradeepa, E. Srinivasana, S. Himavathib, Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten, IJET, Vol. 25, Issue No. 2, 2012.

[15] Amandeep Kaur, Seema Bhagla, Sunil Kumar, Study of Various Character Segmentation Techniques for Handwritten Off-Line Cursive Words, IJASEAT, 2015

[16] R. S. Sabeenian M. Vidhya, Hand written Text to Digital Text Conversion using Radon Transform and Back Propagation Network (RTBPN), International Conference on Advances in Information and Communication Technologies , pp 498-500, 2010.