INTEGRATING OPTICAL CHARACTER RECOGNITION AND MACHINE TRANSLATION OF FIRST INVESTIGATION REPORT TO ENGLISH

[1]Dnyandev Khadapkar, [2]Deepali Raikar

[1]Student, [2]Assistant Professor
[1]Information Technology and Engineering Department,
[1]Goa College Of Engineering, Farmagudi, Ponda-Goa, India

Abstract: This project aims to translate the hand written devnagri scripts of First Investigation Report to English. Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used as a form of data entry from some sort of original paper data source, whether documents, sales receipts, e-mails, or any number of printed records. It is crucial to the computerization of printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used the machine processes such as machine translation, text-to speech and text mining. The algorithm aims at overcoming many problems in long short term memory and recurrent neural network gradient algorithm, Here the each character text is then normalized to bring all the character in uniform size. The goal of our work is to build the application that will demonstrate high performance of these features when classified using back propagation artificial neural network and create an application that will recognize text in a devnagri script and translate it to English. The user will just have to click a picture of a text and will get the translation of text at the click of a button.

***IndexTerms* - Optical Character Recognition, Machine Translation, First Investigation Report, Long-Short Term Memory, Recurrent Neural Network.**

I. INTRODUCTION

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into machine encoded text. It is widely used as a form of data entry from some sort of original paper data source, whether documents, sales receipts, mail, or any number of printed records. It is crucial to the computerization of printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech and text mining.

OCR is a field of research in Pattern recognition, artificial intelligence and computer vision. Early versions of OCR needed to be programmed with images of each character, and worked on one font at a time. ”Intelligent” systems with a high degree of recognition accuracy for most fonts are now common. Some systems are capable of reproducing formatted output that closely approximates the original scanned page including images, columns and other non-textual components. OCR technology can be used to create a translator application. Our project has one more component, translation. Many times, printed text in another language needs to be translated. Our application will make it very convenient as the user will only need to click a photo of the text and will get the translation at the click of a button. A printed text will be translated to English. The following diagram shows the overall flow diagram of the application from the input image to the output translated text. Fig.1 depicts the OCR approach steps.

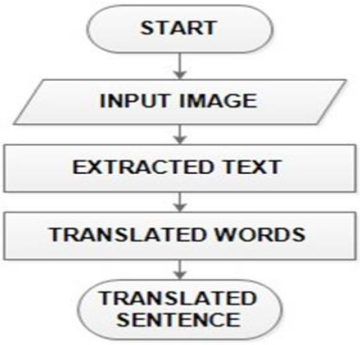


Fig. 1 OCR Approach Steps

II. LITERATURE REVIEW

a) Background

The process of OCR mainly involves 6 phases:

- Acquisition of gray-scale image
- Digitization/Binarization
- Line and Boundary detection
- Feature Extraction
- Feed Forward Neural Network based matching
- Recognition of character based on matching score

The scanned image must be a gray-scale or a binary image, where a binary image is a contrast stretched gray-scale image. That gray-scale image then undergoes digitization. In digitization, a rectangular matrix of 0s and 1s is formed from the image. All RGB values are converted into 0s and 1s (0-black and l-white).The matrix of dots represents two dimensional arrays of bits. Digitization is also called binarization as it converts a gray-scale image into a binary image using adaptive threshold Line and Boundary detection is the process of identifying points in a digital image at which the characters top, bottom, left and right are calculated. Feed Forward Neural Network approach is used to combine all the unique features, which are taken as inputs. One hidden layer is used to integrate and collaborate similar features and if required inputs are adjusted by adding or subtracting weight values. Finally, one output layer is used to find the overall matching score of the network.[1]

b) Analysis of Papers

Bharath V and Shobha Rani [1], The inclination of optical technologies like OCR lies in achieving higher recognition rates with optimal or reduced computational complexities. At present there exist optical technologies for automation of reading the text from document images with almost nearing to 100 percent accuracy. Especially, the Roman language OCR is reliable and robust enough in producing higher accuracies by being able to recognize varying font styles of varying sizes. However for the font style/ size independent OCR one of the main aspect is its computational complexity. It is significant concern to reduce the computational complexities involved in the process of character recognition through a font style / size independent OCR. In this paper, a technique for classification of the font style based on character image is proposed by employing the distance profile features with respect to left, right and diagonal directions of a character image. The major objective of this work is to reduce the complexity of the generic OCR systems by font style recognition. The distance profile features of character images are fed to a support vector machine classifier. For experimentation, the training data sets are comprised of around 10 widely used font styles of upper case letters as well as lower case letters. The testing is conducted with the character images that are extracted from various non-editable document sources comprising of 5 different font styles. The performance of algorithm is found to be satisfactory with an accuracy of 80%.

Sathiapriya Ramiah, Tan Yu Liong, Manoj Jayabalan [2] Smart-phones have been known as most commonly used electronic devices in daily life today. As hardware embedded in smart-phones can perform much more task than traditional phones, the smart-phones are no longer just a communication device but also considered as a powerful computing device which able to capture images, record videos, surf the internet and etc. With advancement of technology, it is possible to apply some techniques to perform text detection and translation. Therefore, an application that allows smart-phones to capture an image and extract the text from it to translate into English and speech it out is no longer a dream. In this study, an Android application is developed by integrating Tesseract OCR engine, Bing translator and phones built-in speech out technology. Final deliverable is tested by various type of target end user from a different language background and concluded that the application benefits many users. By using this application, travellers who visit a foreign country able to understand messages portrayed in different language. Visually impaired users are also able to access important message from a printed text through speech out feature.

Nan Li, Jinying Chen, Huaigu Cao, Bing Zhang, Prem Natarajan [3], While performing testing of our application, it has been noted the OCR developed is very flexible even for non-technical data. The training dataset is an image file saved in suitable (.png or other) format so that it can be used to train the classifier. The model can be scaled for any local language, just by changing training image file and labels in the code.

N Prameela1, P Anjusha1, R Karthik [4], A recurrent neural network is a type of artificial neural network commonly used in speech recognition and natural language processing (NLP). RNNs are designed to recognize a data's constant characteristics and use patterns to predict the next likely scenario. RNNs are used in deep learning and in the development of models that simulate the activity of neurons in the human brain. They are especially powerful in use cases in which context is critical to predicting an outcome and are distinct from other types of artificial neural networks because they use feedback loops to process a sequence of data that informs the final output, which can also be a sequence of data. These feedback loops allow information to persist; the effect is often described as memory.

Tapan Kumar Hazra, Dhirendra pratap singh, Nikunj Daga [5], After converting an image into a gray-scale image, it is analysed and then the difference in spacing and pixel spaces determine the text and its limits so that it may analyse the characters separately without any barrier to the adjust-ability that has been found in this recognizer. Recurrent neural network (RNN) and long short term memory (LSTM) algorithm is used because of its higher accuracy over non-linear multiclass problems.

III. PROPOSED SYSTEM

RNNs are the state of the art algorithm for sequential data and among others used by Apples Siri and Google's Voice Search. This is because it is the first algorithm that remembers its input, due to an internal memory, which makes it perfectly suited for Machine Learning problems that involve sequential data. Recurrent Neural Networks produce predictive results in sequential data that other algorithms can't.

IV. RESEARCH METHODOLOGY

A) Project Methodology

Imagine you have a normal feed-forward neural network and give it the word ’neuron’ as an input and it processes the word character by character. At the time it reaches the character ’r’, it has already forgotten about ’n’, ’e’ and ’u’, which makes it almost impossible for this type of neural network to predict what character would come next.[5] A Recurrent Neural Network is able to remember exactly that, because of its internal memory. It produces output, copies that output and loops it back into the network.

Long-short term memory, LSTM networks are an extension for recurrent neural networks, which basically extends their memory. Therefore it is well suited to learn from important experiences that have very long time lags in between. The units of an LSTM are used as building units for the layers of a RNN, which is then often called an LSTM network. LSTMs enable RNNs to remember their inputs over a long period of time.

B) System Detail Flow

Fig. 2 depicts in detail about working of algorithms on the image to be sent. The process is as follows:

Character Recognition Procedure:

1.) Pre-processing: The pre-processing stage yields clean document in the sense that maximal shape information with maximal compression and minimal noise on normalized image is obtained.

2.) Segmentation: Segmentation is an important stage because the extent one can reach in separation of words, lines or characters directly affects the recognition rate of the script.

3.) Feature extraction: Feature extraction after segmenting the character, extraction of features like height, width, horizontal line, vertical line, and top and bottom detection is done.

4.) Classification: For classification or recognition, back propagation algorithm is used.

5.) Output: Output is saved in form of text format.

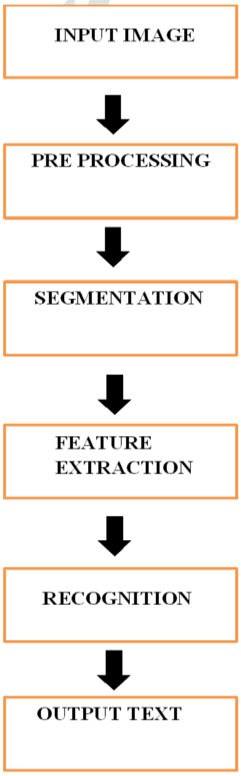
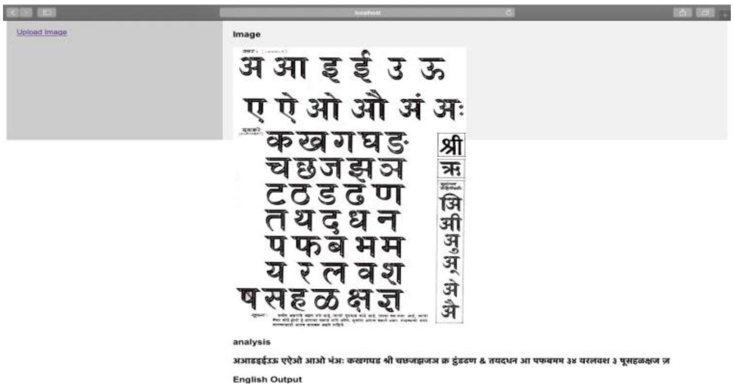


Fig. 2 Character Recognition Steps

V. RESULTS AND DISCUSSION

A) Experiments Conducted

Following is the result when properties like noise reduction, binarization, text Localization, text segmentation, feature extraction, translation filtering technique is applied on scanned background image. Result is also shown in fig. 3, after applying all the above techniques and conversion of gray-scale and inversion of image is done.



analysis

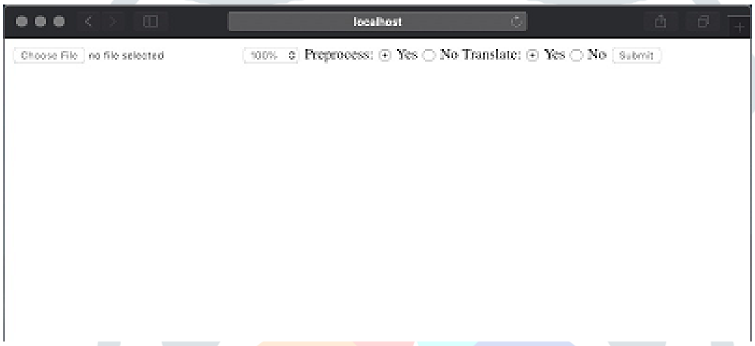
अआइईउऊ एऐओ आओ भंअः कखगघड श्री चछजझजञ क्र ड्रंडढण & तयदधन आ पफबमम ३४ यरलवश ३ षूसहळक्षज ज्ञ

English Output

Fig. 3 Devnagri Script

B) User Interface

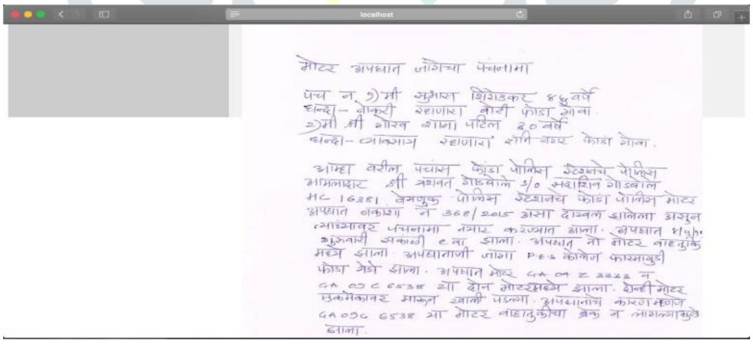
Project has different modules that are choose file, pre-process image, translate and submit as shown in Fig. 4 Localhost as user interface page.



no file selected 100% Preprocess: ● Yes ○ No Translate: ● Yes ○ No

Fig. 4 User Interface Page

Fig. 5 shows the handwritten scanned text image processed with value 120 percentage.



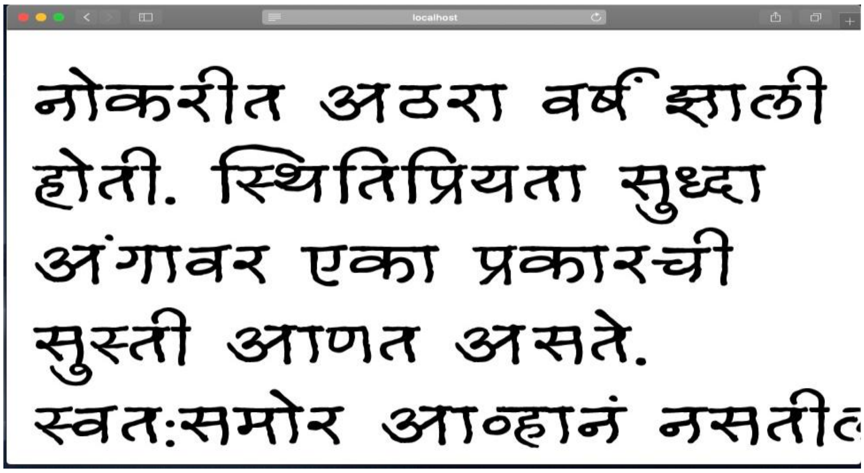
मोटर अपघात जागेचा पंचनामा

पंच नं. १) श्री सुभाष शिरोडकर ४५ वर्षे
धन्दा- नोकरी रहाणार बोरी फोंडा गोवा.
२) श्री श्री गौरव शामा पाटिल ३० वर्षे
धन्दा- व्यवसाय रहाणार शांती नगर फोंडा गोवा.

आम्हा वरील पंचांस फोंडा पोलिस स्टेशनचे पोलिस
अमलदार श्री यशवंत गोडबोले S/o सदाशिव गोडबोले
HC 1638. नेमणुक पोलिस स्टेशनचे फोंडा पोलिस मोटर
अपघात नोंद नं 368/2015 असा दाखल झालेला असुन
त्याच्यावर पंचनामा तयार करण्यात आला. अपघात ४/५/१
शुक्रवारी सकाळी ८ वा झाला. अपघात तो मोटर वाहनांमध्ये
मध्ये झाला. अपघाताची जागा P.E.S कॉलेज फारमागुडी
फोंडा येथे झाला. अपघात मोटर GA 04 E 3323 व
GA 03 C 6538 या दोन मोटरमध्ये झाला. दोन्ही मोटर
एकमेकावर मारून खाली पडल्या. अपघाताचे कारण म्हणजे
GA 03C 6538 या मोटर वाहतुकीचा ब्रेक न लागल्यामुळे
झाला.

Fig. 5 Handwritten FIR Script

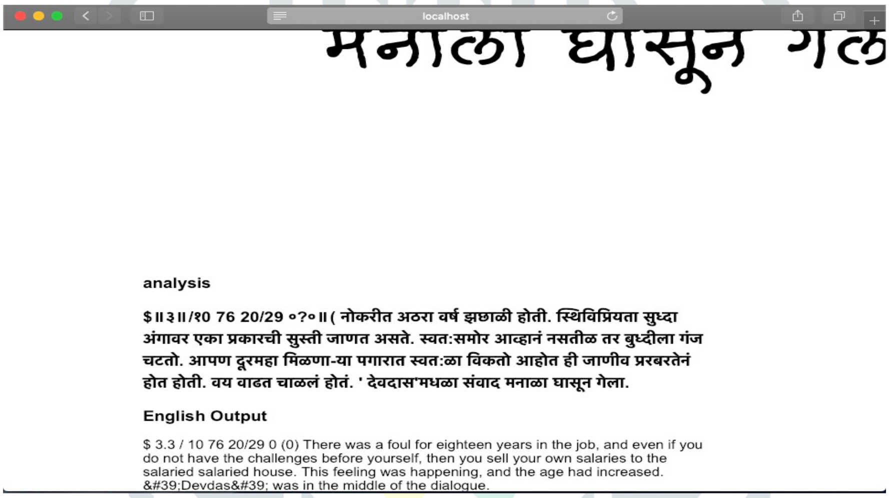
Fig. 6 shows the sample scanned text image with different font is taken and processed.



नोकरीत अठरा वर्ष झाली
होती. स्थितिप्रियता सुध्दा
अंगावर एका प्रकारची
सुस्ती आणत असते.
स्वत:समोर आव्हानं नसती

Fig. 6 Sample Scanned Text Image

Fig. 7, shows depicts the scanned imaged of FIR is processed for the text conversion and translation of devnagri script to English using LSTM and RNN algorithm which detects the text and trains the system using machine learning.



मनाळा घासून गल

analysis

$॥३॥/१0 76 20/29 ०?०॥(नोकरीत अठरा वर्ष झछाळी होती. स्थिविप्रियता सुध्दा अंगावर एका प्रकारची सुस्ती जाणत असते. स्वत:समोर आव्हानं नसतीळ तर बुध्दीला गंज चटतो. आपण दूरमहा मिळणा-या पगारात स्वत:ळा विकतो आहोत ही जाणीव प्ररबरतेनं होत होती. वय वाढत चाळलं होतं. ' देवदास'मधळा संवाद मनाळा घासून गेला.

English Output

$ 3.3 / 10 76 20/29 0 (0) There was a foul for eighteen years in the job, and even if you do not have the challenges before yourself, then you sell your own salaries to the salaried salaried house. This feeling was happening, and the age had increased. &#39;Devdas&#39; was in the middle of the dialogue.

Fig. 7 Final Output Image

IV. CONCLUSION AND FUTURE SCOPE

In this paper, the primary objective is to speed up the process of character recognition in document processing as a result the system can process huge number of documents within less time. The techniques for the recognition of handwritten devnagri text by segmenting and classifying the characters have been proposed in this thesis work. The problems in handwritten devnagri text written by different persons are identified after carefully analyzing the text. To solve these problems new techniques have been developed for segmentation, feature extraction and recognition. A new technique based on header line and base line detection to segment the overlapped lines of text in handwritten devnagri text have been proposed. Determination of the header line is very tough. The position of the header line in particular line of text and header line in a particular word of the same line may vary. The new threshold values for their presence in the word have been proposed. For segmentation of half characters from consonants structural properties of the text are considered. The proposed algorithms are also tested on printed devnagri text and obtained pleasing results. After the segmentation of text, the features are extracted for recognition. A new feature set based on topological features or structural properties of the text have been proposed. A new technique called merging of features for the feature extraction has been proposed in the present work. The overall results obtained with proposed algorithms for segmentation and recognition of handwritten Hindi text is very challenging. The work reported in this thesis can be extended in the following directions. The proposed algorithms used for segmentation of handwritten devnagri text can be extended further for recognition of other scripts. The proposed algorithms of segmentation can be modified further to improve accuracy of segmentation. New features can be added to improve the accuracy of recognition. These algorithms can be tried on large database of handwritten different language text. There is a need to develop the standard database for recognition of handwritten devnagri text. The proposed work can be extended to work on degraded text or broken characters. Recognition of digits in the text, half characters and compound characters can be done to improve the word recognition rate. There is heavy demand for an OCR system which recognizes cursive scripts and manuscripts like Palm Leaves. Speech recognition from OCR, this would help the blind to send and receive information. Moreover speech to text converter through OCR can be worked on.

VI. Acknowledgment

The authors wish to thank everyone who provided insight and expertise that greatly assisted this study. Appreciation to university management for the support and resources provided throughout the research. Also we would like to express our heart-felt gratitude to our beloved principal Dr. Krupashankara M.S for his valuable suggestions and guidance rendered throughout. Also I would like to thank our H.O.D. Dr. Nilesh .B. Fal Dessai and our project coordinator Dr. Aisha Fernandes for providing us with excellent guidance, direction and constant support for successfully performing this project.

References

[1] Bharath V, N. Shobha Rani, A Font style classification system for English OCR.” 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India.

[2] Sathiapriya Ramiah, Tan Yu Liong, Manoj Jayabalan, Detecting Text Based Image With Optical Character Recognition for English Translation and Speech using Android.” 2015 IEEE Student Conference on Research and Development (SCOReD), 2015, Dec, Kuala Lumpur, Malaysia.

[3] B Nan Li, Jinying Chen, Huaigu Cao, Bing Zhang, Prem Natarajan, “Applications of Recurrent Neural Network Language Model in Offline Handwriting Recognition and Word Spotting.” 2014 14th International Conference on Frontiers in Handwriting Recognition, 2014, Dec, Heraklion, Greece.

[4] N Prameela1, P Anjusha1, R Karthik, ”Off-line Telugu Handwritten Characters Recognition using optical character recognition. ”, Online at https://www. http://ijamtes.org/gallery/384-dec.pdf.

[5] Tapan Kumar Hazra, Dhirendra pratap singh, Nikunj Daga, “Optical Character Recognition using KNN on Custom Image Dataset”, 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)”2017,Oct, Bangkok, Thailand.

Authors Profile

Dnyandev Khadapkar, received B.E.(Computer) degree from Padre Conceicao College of Engineering, Verna-Goa, in year 2010 and currently pursuing M.E.(IT) degree from Goa College Of Engineering, Farmagudi, Ponda-Goa in year 2019. His research work includes " Integrating Optical Character Recognition And Machine Translation of First Investigation Report to English".

Ms. Deepali Raikar, received B.E.(IT) degree from Shree Rayeshwar Institute of Engineering & Information Technology, Shiroda-Goa, in year 2005 and received M.E.(IT) degree from Padre Conceicao College of Engineering, Verna-Goa, in year 2011. She is currently the Assistant Professor in Goa College of Engineering, Farmagudi, Ponda-Goa. Her research work includes Operation Research and Compilers.