Multilingual Character Recognition and Character Translation for Indian Document Images

Gunjal Harshada Arun[1] Darekar Swamini Shahaji[2] Rupanavar Divya Bapurao[3] Prof. Pawar A. H.[4]

[1,2,3,4]Department of Information Technology

[1,2,3,4]SVPM's COE Malegaon(bk), Baramati, India

Abstract— Character translation is the technique which is used for the translate character in the one language to another language like Marathi to Hindi or Hindi to English. The translator is taking the character which is segmented by the character segmentation and then translates it into our specific language. The purpose of this project is to develop such a tool which takes an Image as input and extract characters (alphabets, digits, symbols) from it. This Image can be of handwritten document or Printed document. Also it can be used as a form of data entry from printed records. And to translate the recognized character into specified Indian language. In this proposed system detecting text from an image is an important prerequisite for the content based image analysis process. All users are understand the contents of an image or the valuable information, there is need of analysing the text appears in it. The purpose of this design document is to explore the logical view of architecture design, sequence diagram, data flow diagram, and user interface design of the software for programming. The operation such as recognition, detection, segmentation and translation and displaying the text present in the images.

Key words: Character Recognition, Detection, Segmentation, Translation

I. INTRODUCTION

People can be travelling to different places for different reasons. That time it find the difficult communication with these local peoples because traveller people have don't know about this language which is used in that place. They are not understood the words which are written on board. So that time they have need to character recognition and character translation in the different image. Our project will be made the solution of that problem there is image can be capture in camera and recognize character over the image and then it translate the user or people understandable language. So the people can't be create any other problem for the communication or read the any other place information. Character Recognition used in official task in which the large data have to type like post offices, banks, colleges etc., in real life applications where we want to collect some information from text written image.

After the character recognition we can processed on this character that is the character Detection. Character Detection is the technique which is detect the character over the images.

Character Segmentation is the method which is used for the divide the character in multiple form. This method is also used for the character in the number of multiple sub parts for the translation.

Character translation is the technique which is used for translate the character in the one language to another language like Marathi to Hindi or Hindi to English. Translator is take the character which is segmenting by the character segmentation and then translate it in our specific language.

II. OBJECTIVE OF THE SYSTEM

The objective of this research is to study character recognition patterns, character recognition algorithms and tools and eliminate deficiency in identifying character patterns and tools and algorithms for character translation.

- To upload image easy
- To convert the image text in different language.
- To store the multiple image text.

To fulfill this objective some sub objectives were formed which are as following:

1) To use different optical character recognition and translation algorithms.
2) To identify character recognition and translation patterns by Study and analysis of characters for structural and statistical features.
3) To analysis of digits 0 to 9 for structural and statistical features.
4) Collect handwritten data samples and recognize each character and digit using different character recognition tools.
5) To capture different images and recognize each character using character recognition tools (OCR).
6) Identify the common deficiency in most of the character recognition software/tools by calculating the recognition rate of each character and digit and find out the characters and digits whose recognition rate is very less.
7) Designing and development of the model to eliminate the common deficiency identified.
8) For develop the algorithm to implement the above model.
9) Testing and Performance evaluation by analyzing results of model.
10) This online tools is used for the character recognition and character translation.

III. PROPOSED SYSTEM

The country India is a multi-lingual multi-script country and there are twenty two languages. Eleven scripts are used to write these languages and Marathi, Hindi and English are the most popular script in India. First research report on handwritten characters was published in 1977 but not much research work is done after that. In Proposed system, we are working on handwritten Marathi, Hindi and English characters. Many research reports are available towards character recognition but to the best of our knowledge there are not any researches available about multi lingual character recognition and translation.

To get idea of the recognition results of different classifiers and to provide new benchmark for future research,

in this paper a comparative study of Multi-lingual handwritten character recognition results is reported here. To compare the performance, twelve different classifiers and four different features computed from gradient and curvature information of the binary as well as gray scale images are used here. A most commonly used classifier like Support vector machines (SVM), Euclidean distance (ED), Nearest neighbour, k-Nearest neighbour (k-NN), Projection distance (PD), Subspace method (SM), Linear discriminant function (LDF), Modified quadratic discriminant function (MQDF), Mirror image learning (MIL), Modified Projection distance (MPD), Compound projection distance (CPD), and Compound modified quadratic discriminant function (CMQDF) are considered.

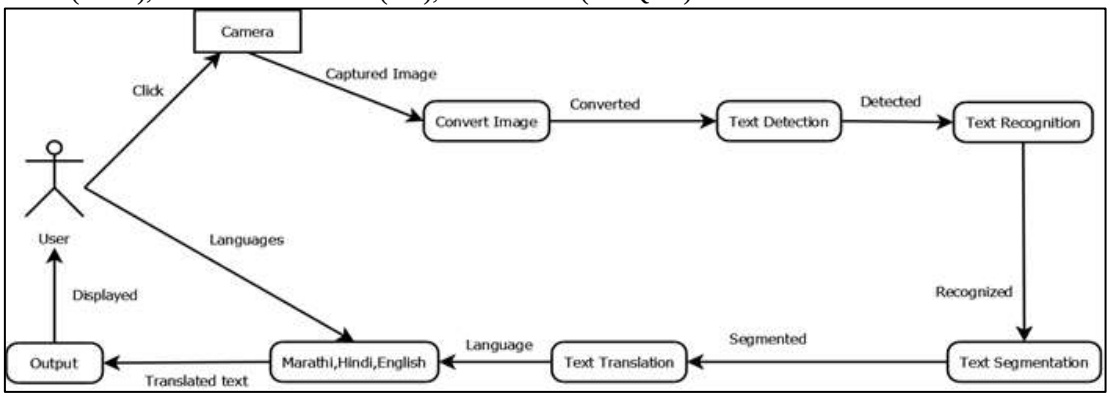

Fig. 1: System Architecture

In the above figure System Architecture, Shows that user can click on the camera for the capture different images that images on the form of Devanagari Script which is converted into other specific languages (Marathi, Hindi, English).Images can converted into binary scale or gray scale.

System can be take a capture image and then process on that images start from the text detection. After the image conversion we can processed on this character that is the character Detection. Character Detection is the technique which is detect the character over the images.

Character Recognition is a process by which text or characters can be input to a computer by providing the computer with an image. The system uses an OCR Engine--a computer program with the specific function of making a guess which letter (recognizable to a computer) an image (recognizable to a human) represents.

Character Segmentation is the method which is used for the divide the character in multiple form. This method is also used for the character in the number of multiple sub parts for the translation.

Character translation is the technique which is used for translate the character in the one language to another language like Marathi to Hindi or Hindi to English. Translator is take the character which is segmenting by the character segmentation and then translate it in our specific language.

IV. Conclusion

The purpose of this system is to develop such a tool which takes an Image as input and extract characters (alphabets, digits, symbols) from it. This Image can be of handwritten document or Printed document. That image can be used as a form of data entry from printed records. And to translate the recognized character into specified Indian language.

References

[1] Parul Sahare and Sanjay B. Dhok, Multilingual Character Segmentation and Recognition Schemes for Indian Document Images. DOI 10.1109/ACCESS.2018.2795104, IEEE Access
[2] Rehman, and T. Saba, Performance analysis of character segmentation approach for cursive script recognition on benchmark database, Digit. Signal Process., vol. 21, no. 3, pp. 486-490, May 2011.
[3] C.Z. Shi, S. Gao, M.T. Liu, C.Z. Qi, C.H. Wang, and B.H. Xiao, Strokedetector and structure based models for character recognition: A comparative study, IEEE Trans. Image Process., vol. 24, no. 12, pp.49524964, Aug. 2015.
[4] M.K. Sharma, and V.P. Dhaka, Segmentation of English o_ine handwritten cursive scripts using afeedforward neural network, Neural Comput. and Applic., vol. 27, no. 5, pp. 13691379, Jul. 2016.
[5] S. Nomura, K. Yamanaka, O. Katai, H. Kawakami, and T. Shiose, A novel adaptive morphological approach for degraded character image segmentation, Pattern. Recogn., vol. 38, no. 11, pp. 19611975, Nov.2005.
[6] V.N.M. Aradhya, G.H. Kumar, and S. Noushath, Multilingual OCR system for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis, Eng. Appl. Artif. Intel., vol. 21, no. 4, pp. 658668, Jun. 2008.
[7] K.C. Santosh, and L. Wendling, Character recognition based on non-linear multi-projection pro_les measure, Front. Comput. Sci., vol. 9, no. 5, pp. 678690, Oct. 2015.

