

Nepali Handwritten Character Recognition System

Santosh Acharya¹, Shashank Dhungel², Ashish Kr. Jha^{3}*

Department of Computer Science and Engineering, Nepal Engineering College, Nepal.

***Corresponding Author**

E-mail Id:-ashishkj@nec.edu.np

ABSTRACT

Even if the technological and digital world is expanding more quickly, there are still many things that are lacking. What a wonderful thing it would be to be able to trust machines to scan any handwritten characters into digital representation. The method for doing this is called optical character recognition (OCR), but there is still much room for improvement. Although there has been work done on it, the technique developed for one language cannot be applied to another due to language variations. Nepali is not a language that is frequently used online. Perhaps this is why there are fewer OCR systems developed using this language. We have made an effort to improve on it so that Nepali characters can be recognized. Basically, the idea is to use a camera to scan Nepali handwriting from hard copy paper, locate the regions in the image where the characters are present, segment those localized parts into characters, and then digitally display each predicted segmented character.

Keywords:-*Optical Character Recognition (OCR), Devanagari handwritten characters, segmentation of handwritten character, training model, Convolutional Neural Network (CNN)*

INTRODUCTION

Nepali Language contains 36 consonants (क, ख, ग, घ, ङ, च, छ, ज, झ, ञ, ट, ठ, ड, ढ, ण, त, थ, द, ध, न, प, फ, ब, भ, म, य, र, ल, व, श, ष, स, ह, क्ष, त्र, ज्ञ) and 10 numerals (०, १, २, ३, ४, ५, ६, ७, ८, ९). They are called basic characters. Vowels can be written as separate letters or with a number of different diacritical marks that are placed above, below, before, or after the consonant they belong to. This type of vowel writing is known as modifying, and the resulting characters are known as conjuncts. Consonants can mix and adopt new shapes on occasion.

Compound characters are the name given to these novel form groupings. When information is scanned through paper documents, the challenge for the present systems is to recognize characters and convert those characters into digital format. As is common knowledge, there are several printed newspapers and books covering a wide range of topics. OCR was designed with those newspapers and books in mind, converting hard copy texts into soft copy texts. The technology analyzes files that are in picture formats like JPEG and PNG. It is exceedingly

challenging to reuse information on hard copy paper and to do line-by-line and word-by-word searches of the papers' contents. OCR software could be useful for searching and reusing the data. There is a tremendous need nowadays for preserving information found in paper documents onto a computer storage drive so that it can be edited or used again in the future.

The potential of data loss in hard copy documentation can be avoided by producing digital copies of the material using the OCR tool. Information may be shared and retrieved from digital data extremely easily. Bank checks, government documents, bill processing software, post code recognition, signature verification, passport readers, and offline document recognition all support handwriting recognition. It distinguishes between individual characters, and the output can be stored in any text format, saving time and effort when manually entering data to be stored in digital format. Using text or data mining, information can be extracted and analysed from websites or web pages.

LITERATURE REVIEW

In 1959, as an early notable attempt in the area of character recognition research was done by

Grimsdale. The origin of a great research work was based on an approach known as analysis-by-synthesis method suggested by Eden in 1968. Eden's work was significant because it explicitly demonstrated a notion that was implied in earlier works: all handwritten characters are generated by a finite number of schematic components. Later, all structural approaches to character recognition adopted this idea.

The paper covers every development in handwritten character recognition in great detail. The kind and quality of the reading material will directly or indirectly affect the most accurate answer that may be given in this area. This study describes a number of character recognition approaches for handwriting recognition systems [1].

A character recognition system's pre-processing procedures are the first stage. The various pre-processing methods used in character recognition systems with various types of pictures are covered in this study. These range from simple handwritten form-based documents to documents with colorful backgrounds and complicated patterns with varying intensities [3].

This article discusses many pre-processing approaches, including skew detection and correction, contrast stretching, binarization, noise reduction, normalization and segmentation, and morphological processing. It was determined that we cannot fully process the image using a single preprocessing method. However, it might not be possible to attain complete accuracy in a preprocessing system even after employing all of the aforementioned techniques [3].

A handwritten character identification system based on the vertical and horizontal line positioning analyzer method was proposed by M. Abdul Rahiman and M. S. Rajasree. The median filter is employed in preprocessing to remove noise. Line and character separation are employed during the segmentation stage to produce isolated characters. Based on line count and location, both horizontally and vertically, the features are extracted. Classification is done using a decision tree

classifier. 91% of people correctly identified people [4].

Using a fuzzy membership function-based technique for HCR, Sumit Saha and T. Som have discussed. To create a standard matrix of a character, ten example images of the character in matrix form are combined. The unidentified character that must be checked for identification is also transformed into an image matrix and compared with each standard matrix in order to be identified using the obtained fuzzy scores [5].

For the classification of Devanagari numerals, Reena Bajaj, Lipika Dey, and Santanu Chaudhary used three different types of characteristics: density features, moment features, and descriptive component features. They achieved 89.6% accuracy for handwritten Devanagari numerals and proposed a multi classifier connectionist architecture to increase recognition reliability[6].

The method for transforming text from a paper document into machine-readable form is described in the publication [7]. Optical Character Recognition is a ground-breaking method that the computer uses to recognize the characters in the document. This paper reviews a number of methods, including OCR employing neural networks and the correlation method.

SVM-based offline handwritten digit recognition was proposed by Renata Neves, Alberto Lopes Filho, Carlos Mello, and Cleber Zanchettin. According to authors, SVM performs better than a multilayer perceptron classifier. Utilizing the NIST SD19 standard dataset, the experiment is run. MLP has the benefit of being able to divide classes that are not linearly separable. MLP, however, is prone to falling into a local minimum zone, where the training will end presuming it has reached the best position on the error surface.

Determining the optimum network design to handle the problem while taking into account the number of layers and perceptrons in each hidden layer is another challenge. A digit recognizer using the MLP structure may not

deliver the expected low error rate as a result of these drawbacks[8].

It has been suggested to use diagonal feature extraction for offline character recognition. It utilizes the ANN model. This neural network recognition system is created by combining two methods using 54 and 69 features, respectively. The neural network recognition

system is trained using both the horizontal and vertical feature extraction methods in order to compare the recognition effectiveness of the suggested diagonal feature extraction approach. The recognition accuracy produced by the diagonal approach of feature extraction is determined to be 97.8% for 54 features and 98.5% for 69 features [9].

METHODOLOGY

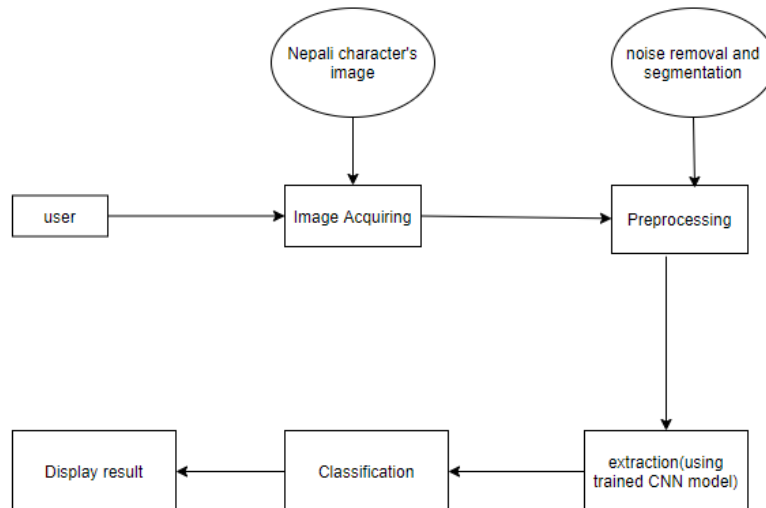


Fig.1:-System Architecture

Nepali Handwritten Recognition System consists of different components that make up the complete functionality of the system.

- User: Person who uses or operates system.
- Image Acquiring: Nepali Character's image is acquired by the NHWCR system through a source device for further processing.
- Pre-processing: Here, the image is acquired from above component, pre-processing tasks like noise removal, dikka removal and segmentation are performed on acquired image to make it feasible for CNN model.
- Feature Extraction: CNN model is used here for feature extraction from the pre-processed image. The model is trained with the training data set which contains many samples of Nepali characters.
- Classification: Here, the system classifies the input character among the characters from क to ख and from ० to ९.

- Display result: Finally, the classified character is shown as the recognized output character.

A raw image of Devanagari character is taken as input. And many pre-processing modules are implemented on the image to make it eligible for the CNN model.

Data Cleaning

The practice of correcting or deleting inaccurate, damaged, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. There are various ways for data to be duplicated or incorrectly labeled when merging several data sources.

Even if results and algorithms appear to be correct, they are unreliable if the data is inaccurate. Because the procedures will differ from dataset to dataset, there is no one definitive way to specify the precise phases in the data cleaning process. But it is essential to create a template for your data cleaning procedure so you can be sure you are carrying it out correctly each time.

Segmentation

The region of the image where the character is located should be determined which is the region of our interest. So, the following function helps to determine the region of interest. Since our image will be in vales of 0's and 1's, this function takes the region which has 1 value.

This process is known as segmentation of image on the basis of their values. The segmentation occurs in 3 steps:

After word segmentation dikka (horizontal line on top of the characters) should be removed so that character can be segmented. Here segmentation process is computed by inheriting the ROI function.

Line Segmentation

The initial stage in separating the text from the text document is line segmentation. This line

segmentation must be done from top to bottom and from left to right. Either the base line or the head line is used to segment the lines.

Word Division

Words from the split lines will then be extracted. Based on space, these words are divided. Vertical lining on the gap that exists between two words can be used to segment the text.

Removal of Dikka

Word segmentation is followed by dikka elimination. The longest line is eliminated together with the horizontal straight line that is found at the top of the word.

Character Division

The ultimate stage of text segmentation is character level segmentation. Segmented words will be used to segment characters.

RESULTS AND FINDINGS

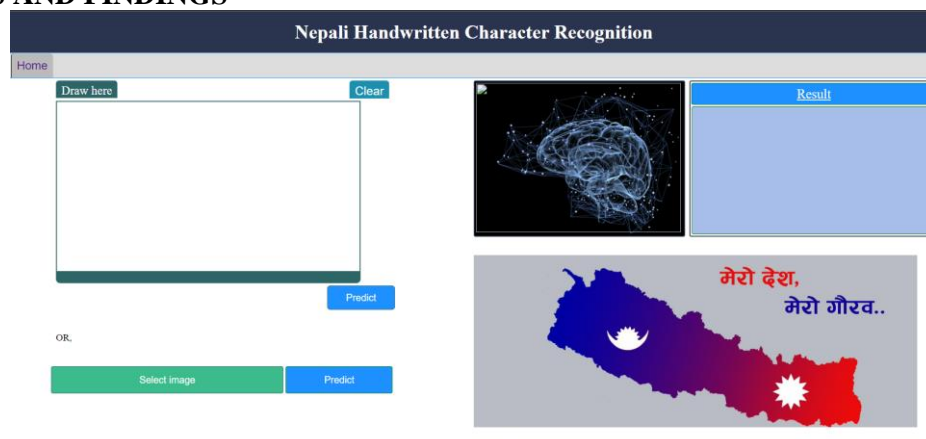


Fig.2:-Image of Created Module



Fig.3:-Testing Data from Image

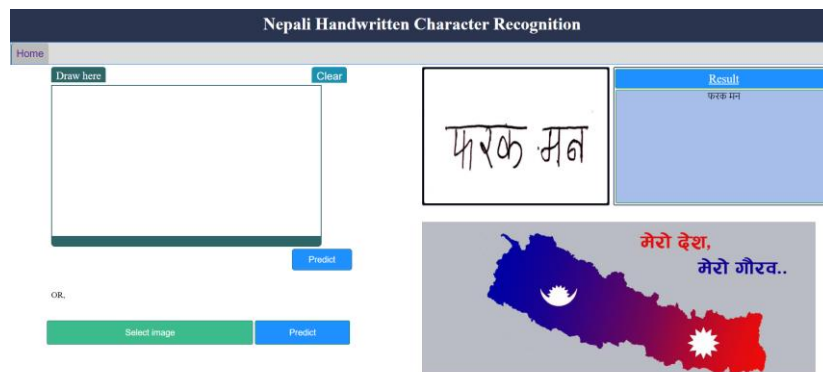


Fig.4:-Result of Previous image detected

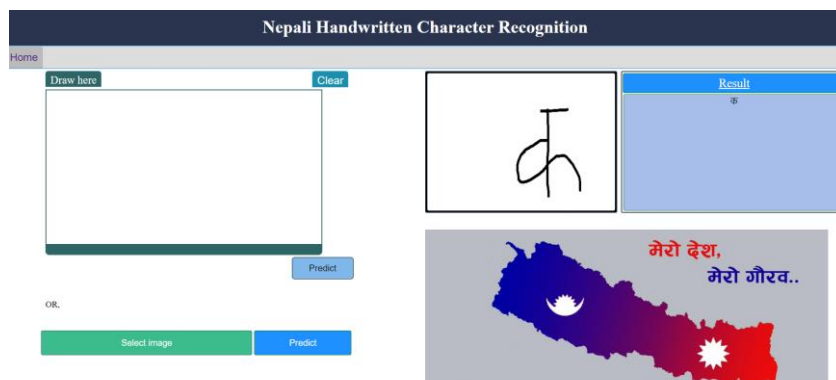


Fig.5:-Image of live data testing

Test case: Model Accuracy Test

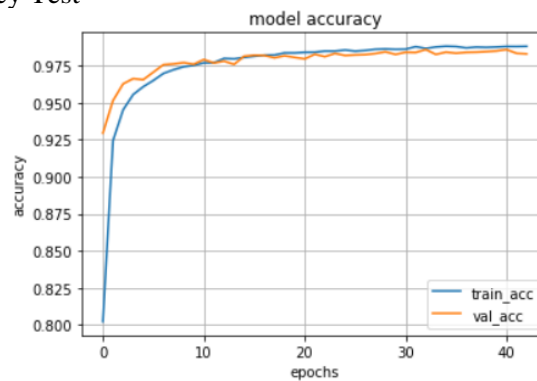


Fig.6:- Model Accuracy Testing

Test Case: Model loss Testing

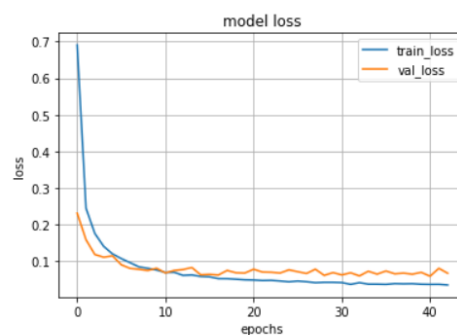


Fig.7:-Model of Loss Testing

Here, accuracy graph shows the accuracy of the model is 0.9816 on testing dataset and 0.9859 on training dataset thus it concludes the efficiency of the model is 98.16%.

CONCLUSION

Nepali characters are read by the system in the web-based Handwritten Character Recognition System, which then displays the character that was recognized using the CNN algorithm. The Convolutional Neural Network (CNN) technique was employed in this system to extract features from the pre-processed image of the Nepali character, which enabled the system to identify that character. In order to recognize handwritten characters when working on the NHWCR system, the complexity of noise removal techniques, segmentation techniques, and other pre-processing tasks were explored. Additionally, the technology was able to transform handwritten Nepali text on paper into an editable digital format.

REFERENCES

1. Purohit, A., & Chauhan, S. S. (2016). A literature survey on handwritten character recognition. *IJCSIT International Journal of Computer Science and Information Technologies*, 7(1), 1-5.
2. Acharya, S., Pant, A. K., & Gyawali, P. K. (2015, December). Deep learning based large scale handwritten Devanagari character recognition. In *2015 9th International conference on software, knowledge, information management and applications (SKIMA)* (pp. 1-6). IEEE.
3. Kumar, G., Bhatia, P. K., & Banger, I. (2013). Analytical review of preprocessing techniques for offline handwritten character recognition. *International Journal of Advances in Engineering Sciences*, 3(3), 14-22..
4. Rahiman, M. A., Rajasree, M. S., Masha, N., Rema, M., Meenakshi, R., & Kumar, G. M. (2011, April). Recognition of handwritten Malayalam characters using vertical & horizontal line positional analyzer algorithm. In *2011 3rd International Conference on Electronics Computer Technology* (Vol. 2, pp. 268-274). IEEE..
5. Saha, S., & Som, T. (2011). Handwritten character recognition using fuzzy membership function. *IJETSE International Journal of Emerging Technologies in Sciences and Engineering*, 5(2).
6. Bajaj, R., Dey, L., & Chaudhury, S. (2002). Devnagari numeral recognition by combining decision of multiple connectionist classifiers. *Sadhana*, 27(1), 59-72..
7. Charles, P. K., Harish, V., Swathi, M., & Deepthi, C. H. (2012). A review on the various techniques used for optical character recognition. *International Journal of Engineering Research and Applications*, 2(1), 659-662.
8. Neves, R. F., Lopes Filho, A. N., Mello, C. A., & Zanchettin, C. (2011, October). A SVM based off-line handwritten digit recognizer. In *2011 IEEE international conference on systems, man, and cybernetics* (pp. 510-515). IEEE..
9. Pradeep, J., Srinivasan, E., & Himavathi, S. (2011, April). Diagonal based feature extraction for handwritten character recognition system using neural network. In *2011 3rd international conference on electronics computer technology* (Vol. 4, pp. 364-368). IEEE.