CNN BASED CHARACTER RECOGNITION FOR ISOLATED HANDWRITTEN GUJARATI CHARACTERS AND NUMERALS

Devanshi Shah
Information Technology
C S Patel Institute of Technology,CHARUSAT
Changa, Gujarat, India
20IT128@charusat.edu.in

Heli Shah
Information Technology
C S Patel Institute of Technology,CHARUSAT
Changa, Gujarat, India
20IT130@charusat.edu.in

Prof. Sanket Suthar
Information Technology
C S Patel Institute of Technology,CHARUSAT
Changa, Gujarat, India
sanketsuthar.it@charusat.ac.in

***Abstract* - The necessity of recognizing the handwritten characters is increasing day by day because of various application and increase in technology. Objective of our project is to provide the effectiveness and efficient way to recognize the Gujarati handwritten isolated characters. Here we have worked on box extraction code with the help of python to extract each character from single boxes from whole image and we have extended (CNN) convolutional neural network for recognizing Gujarati handwritten characters. With the help of CNN, we will test the accuracy of 10 classes for digits, 12 classes for vowels and 35 classes for consonants.**

Key Terms – Gujarati Character Recognition, Handwritten Character Recognition, Segmentation, Image Classification.

INTRODUCTION

In the current situation the importance of handwritten character recognition is increasing and application is prevalent and technology vision too. With the improvement in the technology government is also trying to computerized the information repository which includes a very huge amount of handwritten script. Written script is an old method which requires a very huge number of man - power and it requires a huge amount of time too. Handwritten character recognition is an automate process and this automation with help in many areas such as converting the text of hard copy to a soft copy, signature verification, handwritten licence recognition and etc. But it is very challenging task because collection of handwritten characters is comparatively difficult with respect to printed characters. Around 55 million people speak Gujarati language. Gujarati handwritten characters have versatility in size, shape, thickness, noisiness. strokes and writing style of different age group people with different bold of pens. Therefore, we are trying to propose a sophisticated model like CNN to extract features from images automatically without any explicit description. Some similar characters differ from one another by just single dot mark or a line mark. This feature extraction task becomes a lot of challenge because of different writing styles with different strokes and different spacing. And these are two factors of lose in the accuracy of characters and numerals.

In this project we are intending to solve the problem of explicit feature extraction and a propose a method that will automatically select and extract feature from the character images of different styles and spacing. We have considered 35 basic letters, 12 modifiers and 10 numerals and we are trying to propose a method that will have a better accuracy than the current existing methods.

The basic characters of the Gujarati language are shown in Table 1. In addition to basic characters, it contains about 420 compound characters which are formed by one or more basic characters. The formation mechanisms of these compound characters are shown in table 2. And table 3 shows the handwritten words in Gujarati.

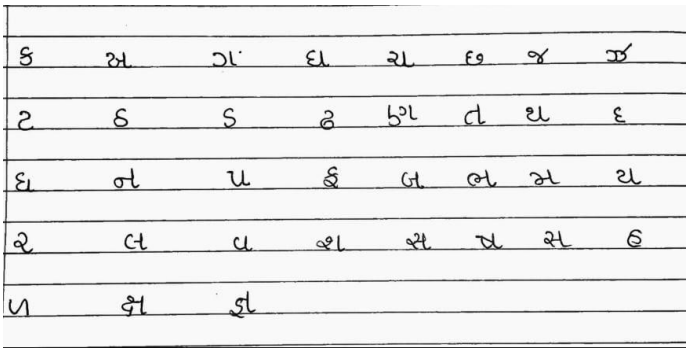


Table 1

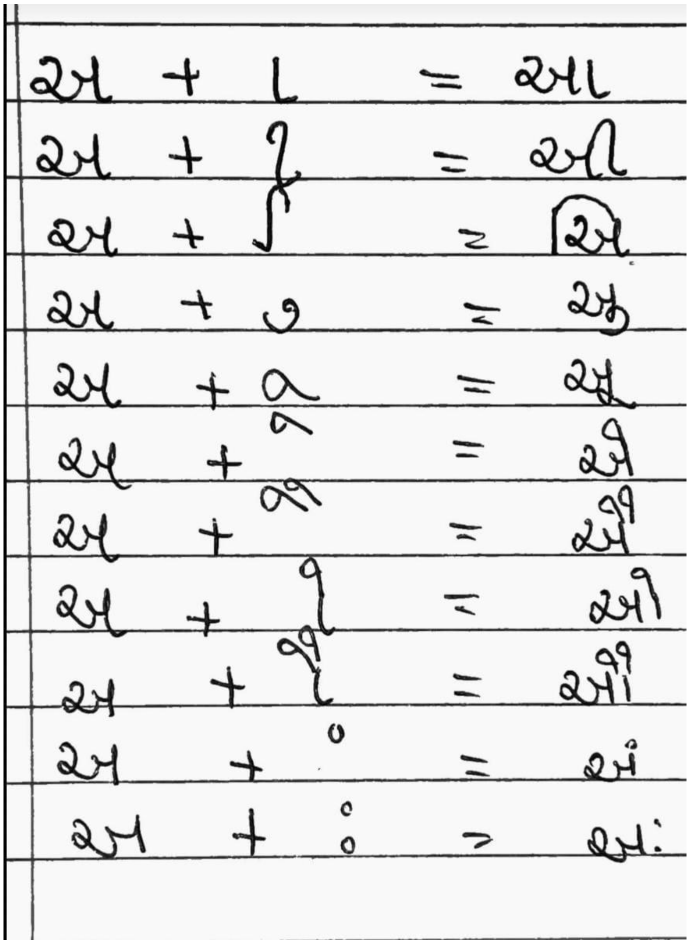


Table 2

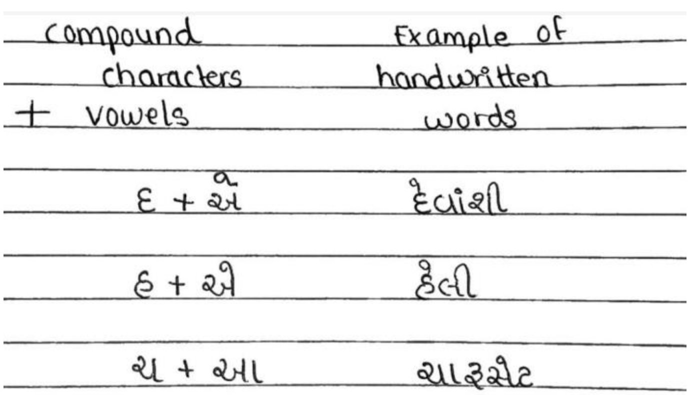


Table 3

REASON TO LEAD THIS PROJECT:

There are many monuments, places to visit in Gujarat state. But tourists might be facing the language barrier. To overcome this problem, we are leading with this project. And there would be quiet easy task to work on printed characters because we won't be able to find more than 60-70 variations in printed characters but in handwritten, if we collect datasets from smaller age group to bigger, we can get lots of variations. So, this is what are basic reason to lead with this project.

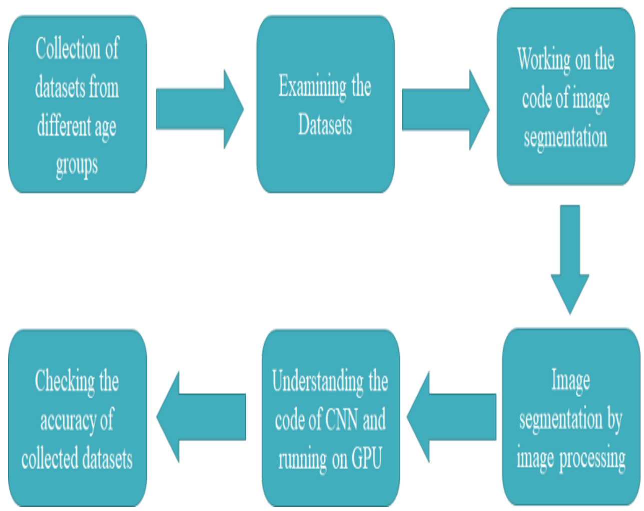
SURVEY

There are many projects done based on the type we are working on. But we weren't able to find such a project in our regional language i.e., Gujarati. So, we are trying to work on this project using Guajarati language.

SCOPE

The scope of this project is for the beneficial purpose for all tourist people to convert the language Guajarati in particular selected language.

FLOWCHART OF THE PROJECT



METHODOLOGY:

Software Requirement: PyCharm (64-bit)

Hardware Requirement: 8 GB RAM laptop

i5/i7 processor

Scanner for scanning

Datasets with variations

GPU for checking accuracy of datasets through CNN code

Language requirement: Python and some concepts related to OpenCV, neural networking and machine learning.

THE FUNCTIONALITY

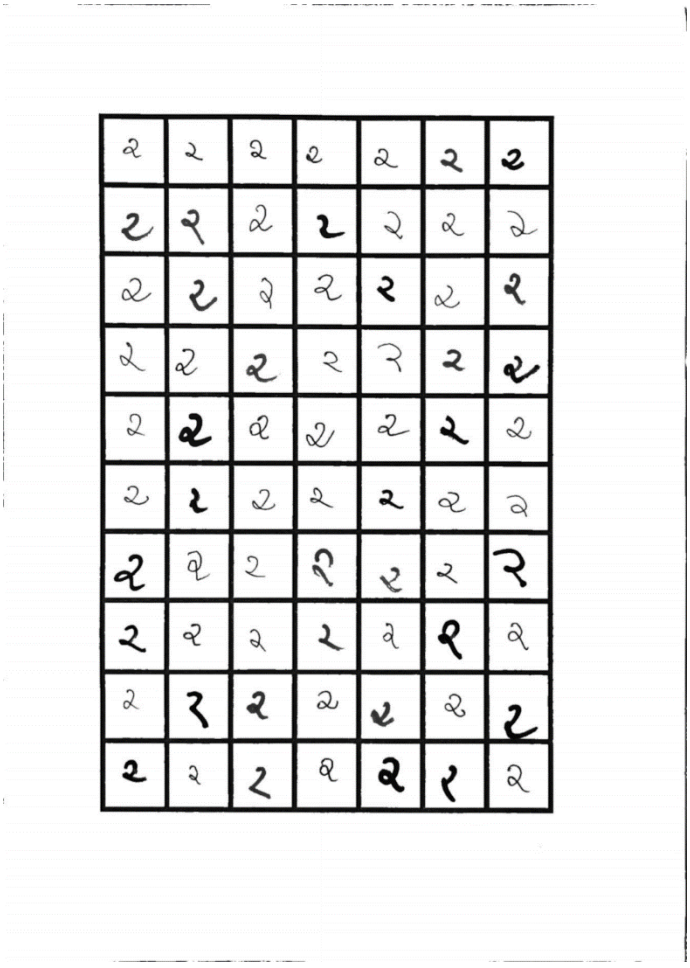
Datasets: The collection of data is called dataset. Datasets are collection of related sets of information that is composed of separate elements but can be manipulated as a unity by a computer.

Python code for box detection: Box detection is a python package based on OpenCV which allows us to detect rectangular type shapes easily. Main purpose of this library is to provide helpful functions for processing document images for various applications.

CNN code: A convolutional neural network (CNN) is a neural network that has one or more convolutional layers which are used for image processing, classification. Segmentation, etc. CNN is a kind of filter over the input. CNN is a type of Artificial neural network used in image recognition and processing which is specified for particular fixed pixels.

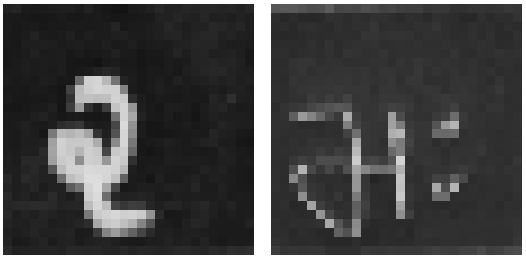
RESULTS

INPUT



Output

Extracted handwritten characters



FUTURE WORK:

We are trying to resolve the error in the CNN code of Devanagari later on we will check the accuracy of datasets through that code and then we would be working on CNN code for gujarati language.

ACKNOWLDEGEMENT

I would like to express my special thanks of gratitude to my Sir, Sanket Sir who gave me the golden opportunity to do this wonderful project on the topic image segmentation and character recognition, which also helped me in doing a lot of Research and I came to know about so many new things.

REFERENCES

[1]https://www.researchgate.net/publication/49592833_A_Review_of_Research_on_Devnagari_Character_Recognition

[2]https://www.researchgate.net/publication/350489483_Bangla_Handwritten_Character_Recognition_Using_Extended_Convolutional_Neural_Network
https://youtu.be/2-Ol7ZB0MmU