TELUGU OPTICAL CHARACTER RECOGNITION USING CLOUD COMPUTING AND PYTHON

MAHESHA C

Department of computer Application Jain University Bangalore, India

ABSTRACT

This is the generic method of cloud based Optical Character Recognition(OCR). Where individuals can extract the actual Copyable content and text from pictures which has just unwritable contents. Optical Character Recognition(OCR) is the early technique based and Artificial Intelligence and Machine Learning(AI/ML) where many authors used this with neural network to extract the required data from unstructured documents. The major approach of this work is to help the individuals who were persistently typing the substance from meaningful pictures to get the Text from it. Here in this project I have utilized the Bot where we have to transfer the picture based electronic records or text based images and in the backend the picture will be extracted into organized writable information. In data entry job field many individuals have to type more by watching the unstructured text from pictures to help the individuals and the needy people we proposed the system called OCR bot using Telegram. By using telegram people can send the images and get the content what is there in that image. This assistance may accommodate for all understudies and representatives to save time simply by extracting rather than typing each letter.

KEYWORDS: Optical Character Recognition, Artificial Intelligence, Cloud computing, Machine Learning

Received: Mar 03, 2021; **Accepted:** Mar 23, 2021; **Published:** May 10, 2021; **Paper Id.:** IJCSSEITRJUN20219

INTRODUCTION

As we know most of the people busy all over the day especially in typing, it is the biggest problem that so many are facing nowadays individuals has no patience for typing lines and lines of text by watching the printed notes.

According to the survey conducted by Emma goldsmith, 73% of keyboard users facing the hand pain and 30% of people who are using computer keyboard are facing the same hand pain difficulty [1]. Like this so many students facing typing troubles to reduce these problems we come up with OCR bot.

In this proposed system, we built a telegram bot which can satisfy individuals who are looking for text extraction from images they can upload a picture containing text and get the text from telegram. Basically OCR bot is used for extraction of handwritten and typed hardcopy to editable data easily. I have used python and cloud computing for API web hook calls to achieve this project and telegram Bot father and Bot token for communication [2].

Here in this framework I proposed a communication bot where individuals can upload and get the structured and copy able data which will beat the existing system disadvantages like high waiting time and lengthy process mismatched extraction, web operations.

To overcome these problems everyone can use our OCR bot, people need not to do lengthy process and wait for the backend process or mismatch extraction they can simply just by uploading the images can get the

standard electronic data from it [3]. And the major benefit of my proposed system is that the printed document can be extracted into any other languages too in the area of artificial intelligence(AI) and image processing has more attention on Optical character recognition which was most popular in 1900s to extract the data from printable text to standard text [4].

No cost estimation and time estimation as well as man power needed for this proposed system. It is completely Open source.

Motivation

The major motivation of this project is to assist the students or professionals who are looking to save the time by not typing each line by line of text from printed documents. We can observe many are facing stress and hand pains by keep typing all over the day to overcome these problems and serve the needy people we come up with the solution called optical character recognition using cloud computing and python and machine learning.

BACKGROUND

In the proposed system I am using Cloud computing and python where I code the program and to deploy that into telegram bot by using token given by bot father [5]. And cloud for web hook call from python the extraction process will be done in cloud

User need to upload the picture containing text to the telegram bot and the bot will pass the image to python and from python the image will be transmitted to cloud the tesseract will extract each letter by scanning the picture and after processing the result will be sent back to telegram bot through python.

REQUIREMENTS

Hardware Requirements:

- 2GB or higher RAM recommended (minimum 1GB)
- Processer: core 2 duo or higher version
- *Software requirements:*
- Python3
- Pytesseract
- Telegram
- Microsoft windows 10/8/7(32bit or 64bit)

PROPOSED SYSTEM

In this proposed system, the diagram of how optical character recognizer will work we will get to know and the methods we have used here we can briefly can get [6].

Cloud computing provides different kinds of web hook and many fulfilments which we can use for getting the process done there and the result has to be generated using cloud computing. Here in this project I am using Cloudmersive to call and the extraction process will be done in cloudmersive.

For the workflow here I have used python for getting the input from the user and to send the input data to cloudmersive and extraction process and getting the result information back from cloudmersive and sending the output to the user. The whole process goes with API key, which was generated by cloudmersive [7]. which we have placed in python

and also telegram bot key.

Overview of Architecture

Here in this figure 1 describes that the user will upload the picture, which has Telugu text in it so the telegram API bot will send the query to python as we have linked our Telegram bot key with python so python will send the query to cloudmersive using API key provided by cloud.

And the result will be extracted from cloud and send back to the python then using telegram API the result will be served to telegram.

Figure 1: Architecture of Optical Character Recognizer

Here in cloudmersive.com the picture will be extracted and sent back to telegram using API key using the workspace of python the whole input and output are processed in Python [8].

In this project, I have used other method called tesseract for taking the text from the image we upload so the user will get the text in Telugu language and users can copy the text from the output generated by tesseract.

Tesseract is the function where people commonly used in for extracting the text from images in windows or in Linux and here, I have used python and installed the tesseract library and some tools like pytesseract and we should install required languages like Telugu, Tamil, Kannada, English etc.

Input Analysis

Telegram users will send the picture containing telugu text in it and the pictures will be tokenized in python and using Nano technology the text in the picture will be extracted bit by bit and it will match with the language and the output will be generated by removing stop words with NLTK corpus and the response will be served to the users who sent the picture [9].

For example if the picture having the text of hello and the bits will be compared using Nano technology and machine learning after comparing the dark and light texts the text will be compared with alphabets so the result will be displayed to the users.

Figure 2: Input Analysis

Coding Methodologies and Training

For coding part we must have to install python and visual studio code where we can code for python and also we have to install tesseract with all available and required languages so that the python should be ready to extract and connect with both the telegram and cloudmersive both and also getting and giving the inputs and outputs to the users.

To develop the bot we have to set the query set what and all the other doubts can user's type in telegram and what will be the output for their quires all we have to provide in python as a input and output [10]. So the intents and entities are created as a queries and responses to that using coding in python.

Figure 3: Development and Training

Optical Character Recognition and Response

In this project I am using these tools and libraries

- Cloudmersive
 - Cloudmersive.com is the private cloud platform used for Optical character recognition API calls
 - To import the text from uploaded picture
- pytesseract
 - This is the python library which is used for non-cloud extracting just by installing library.
 - This library uses to translate and extract
- Telegram
 - Telegram is for users who need to extract the text from any images which has text in it.

- For connection we will use Botfather API key.

Here I am using Cloudmersive to extract Telugu text and give copyable text in telugu so people and students who need to type more can easily get the text from that picture or readable images which are uploaded to telegram bot within seconds.

User will send the picture to the telegram bot we have created and the Bot API has linked with python to interact. Also API key has generated in Cloudmersive and or we can use tesseract to extract [11] [12]. Google tesseract is a software or tool used for optical character recognition and so many languages can use this tool to extract text from the given picture [13].

Figure 4: Request & Response Processing using Tesseract OCR

To build the connection to communicate with cloudmersive we have to register with cloudmersive.com and need to provide our credit card for that. After signup it will provide API key for us using that API key we have to build the connection with python to communicate and interact with it. So the connection will be established [14].

RESULT AND DISCUSSION

Here I explained about how to communicate with botfather and how to get the token for our bot and also the process of connection and communication all discussed here. And the result of the project.

Bot setup for Communication

Telegram provides the Application programming interface which is used for external api calls and to create bots and bot tokens for communication. To fulfil the user needs we come up with this idea of bot responses so here in this section I have started a botfather and requested a new bot which will be unique and the name of the bot should be end with “bot” to create this bot we no need to verify our application just by using bot we have to create a bot for us. By using the token generated by botfather we have to connect with our project where we can able to communicate with the bot.

Figure 5: telegram Bot-Father Wizard

RESULT

Initial Response

To welcome the telegram bot users I made little greeting type of response following with the users input, so that the user will get better response and good greeting to begin with and to start the extraction process I have said to send the picture in the next line these all responses followed by the /start command initiated by the user. The response will mention the user's username.

Figure 6: initial response

Other Random Text

When user type some random text the bot will reply to the user as the text what the user told and what is the next step you have to do in that process. To continue the conversation I have made this response and the response will be satisfy to the user. This response containing the username and the message what the user sent to us and also letting the user know the next step as to upload the picture.

Figure 7: random Text

start /help /donate Commands

Here I have initialized some basic commands these commands are /start which is to begin the conversation and /help command is to let the users know the details process and the contact details. Like the user uses the service and then they can able to donate to my UPI id or if the user face any issue they can contact us by clicking the /help command and they can get our mail id to communicate through

Figure 8: Commands

These commands are for users need, the required field will deliver the response to the user accordingly, and the next step is extraction process.

Text extraction

Here is the main and important part of our project is text extraction of Telugu text. Here I have tried many test trails by uploading Telugu text images to the telegram bot whenever user uploads a picture cloumersive will redirect and extract the text and sent back to the telegram via python.

Figure 9: Text extraction

In the text extraction system I designed to do only Telugu text extraction, we can do any language extraction we want, and we can copy the text easily. No need of typing the whole text and wasting the time.

CONCLUSIONS

Optical character recognition process helps the people who has no time to type each letter from an image or document. Everyone who uses telegram can adopt this opportunity and can use this bot for text extraction. In this model, we have created the bot in that the user will send the picture containing Telugu text that text will be compared with the pre-built text in a tesseract and the most accurate result will be posted as an output.

To build the system I have used cloudmersive, which is the cloud based text extraction site where the input picture

will be extracted as a readable and digital text. In addition, python for connecting cloudmersive and telegram by using API key and the bot token address so that whenever we run the code the process of extraction done if the user send the valid Telugu text image to the telegram.

Future work is just to include multiple languages into it so that it may helpful for the people who need to extract the text in English Tamil French etc.

REFERENCES

- [1] B. Maharrey, "A Neural Network Implementation of Optical Character Recognition," *IEEE*, no. CSSE10-05, 2009.
- [2] A. M. Sabu and A. S. Das, "A Survey on various Optical Character," *IEEE Conference on Emerging Devices and Smart Systems*, pp. 152-155, 2018.
- [3] C. Liyanage, T. Nadugodage and R. Weerasinghe, "Developing a commercial grade Tamil OCR for recognizing font and size independent text," *International conference on Advances in ICT for emerging regions*, 2015.
- [4] R. Parthiban, R. Ezhilarasi and D. Saravanan, "Optical Character Recognition for English Handwritten Text Using Recurrent Neural Network," *IEEE xplore*, 2020.
- [5] M. A. Hasnat, M. R. Chowdhury and M. Khan, "An open source Tesseract based Optical Character Recognizer for Bangla script," *International Conference on Document Analysis and Recognition*, Dhaka, 2009.
- [6] N. Prameela, P. Anjusha and R. Karthik, "Off-line Telugu Handwritten Characters Recognition using optical character recognition," *International Conference on Electronics, Communication and Aerospace Technology*, pp. 223-226, 2017.
- [7] Chang and I. W. a. Hsiao-Chen, "Signboard Optical Character Recognition," *china*, 2019.
- [8] S. Garg, K. K. Gupta and N. Prabhakar, "Optical Character Recognition using Artificial Intelligence," *International Journal of Computer Applications*, vol. 179, no. No.31, pp. 14-20, 2018.
- [9] R. Deepa and K. N. Lalwani, "Image Classification and Text Extraction using Machine Learning," *Third International Conference on Electronics Communication and Aerospace Technology*, no. 45616, pp. 680-684, 2019.
- [10] Q. Li, W. An, A. Zhou and L. Ma, "Recognition of Offline Handwritten Chinese Characters Using the Tesseract Open Source OCR Engine," *8th International Conference on Intelligent Human-Machine Systems and Cybernetics*, pp. 452-456, 2016.
- [11] H. Lu, B. Guo, J. Liu and X. Yan, "A Shadow Removal Method for Tesseract Text," *IEEE xplore*, Beijing, 2017.
- [12] M. Shen and H. Lei, "Improving OCR Performance with Background Image Elimination," *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 1566-1570, 2015.
- [13] S. K. Singla and R. K. Yadav, "Optical Character Recognition Based Speech Synthesis System Using LabVIEW," *Journal of Applied Research and Technology*, vol. 12, pp. 919-926, 2014.
- [14] H. Sidhwa, S. Kulshrestha, S. Malhotra and S. Virmani, "Text Extraction from Bills and Invoices," *International Conference on Advances in Computing, Communication Control and Networking*, pp. 564-568, 2018.