

Optical Character Recognition

SABER SHEYBANI¹ AND SUSHMITA SIVAPRASAD¹

¹School of Informatics and Computing, Bloomington, IN 47408, U.S.A.

* Corresponding authors: sheybani@umail.iu.edu, sushsiva@umail.iu.edu

project-000, April 9, 2017

Optical Character Recognition is a technology for converting images into machine encoded text format. In this project, the input data is in PNG format and our goal is to recognize the words/letters in the image as accurately as possible and convert the dataset into TXT format. The heart of OCR is a classification algorithm which will be implemented using Python programming language. The algorithm will be deployed using the Ansible technology [1] to remote virtual clusters.

© 2017 <https://creativecommons.org/licenses/>. The authors verify that the text is not plagiarized.

Keywords: OCR,ansible,classification

<https://github.com/SushmitaSivaprasad/sp17-i524/tree/master/project/S17-IR-P012/report/report.pdf>

1. 1.INTRODUCTION

This project proposal provides an overview on how we plan on implementing the OCR technology. It gives a background on the kind of technology that has been used, delving into some of the basic concepts used in the implementation process. We have also discussed some important applications of this technology in the real world.

2. 2.BACKGROUND

2.1. 2.1 OCR Technology

Optical Character Recognition is a technology which is used to convert different types of documents that can be in the form of scanned papers (raster images) or PDF into an editable and searchable form [2]. The images can be in either the basic black & white or multicolored. The technology first analyzes the structure of the document and divides it into smaller segments. Finally, individual characters are singled out one by one and fed to a classification algorithm which will return the closest letter that the individual character could possibly be identified with.

2.2. 2.2 Ansible

Ansible is an IT automation tool. It uses YAML in order to issue the state of the server [2]. Ansible implements the internal command that is required to reach that state which depends on the operating system. The ansible playbook which consists of these internal commands can be applied across any server or service. There is no requirement to install an additional software on the target system as the commands are run over an SSH session.

2.3. 2.3 Feed Forward Neural Networks

Artificial Neural Network is a paradigm in computing, inspired by the structure of biological nervous systems. It consists of a network of processing units, where the output of each unit is a nonlinear function of its weighted inputs that come from other units. Such network can be trained to solve different kinds of problems, including classification and clustering. A feed forward neural networks is one in which the neurons are organized in a number of layers and each layer only feeds to the next one, but not to the previous one (no feedback). However, in a back-propagation process, the errors from one iteration of classification will be fed back from the output to the network, in order to modify and improve the network for next iterations.

3. 3.ANSIBLE DEPLOYMENT

We will be using Ansible [1] for running the OCR algorithm. The jobs will be collected and organized in a Playbook [3] and run on virtual clusters provided by Chameleon Cloud [4]. The tasks will include Installing the essential libraries on the remote machine and running the program.

4. 4.OCR IMPLEMENTATION

Optical Character Recognition have already been developed in numerous ways, focusing on different goals. For our purpose, various classification algorithms such as K-Nearest Neighbor and Neural Network (multilayer perceptron) can be used. For this project, a feedforward, back-propagation Neural Network will be used. The steps are as follows: Preprocessing: The input images need to be segmented into units that each of them keep only one glyph (symbol). Also, the colored or grayscale images will be binarized. Feature extraction: The glyphs will be decomposed into features like lines, closed loops, line direction, and

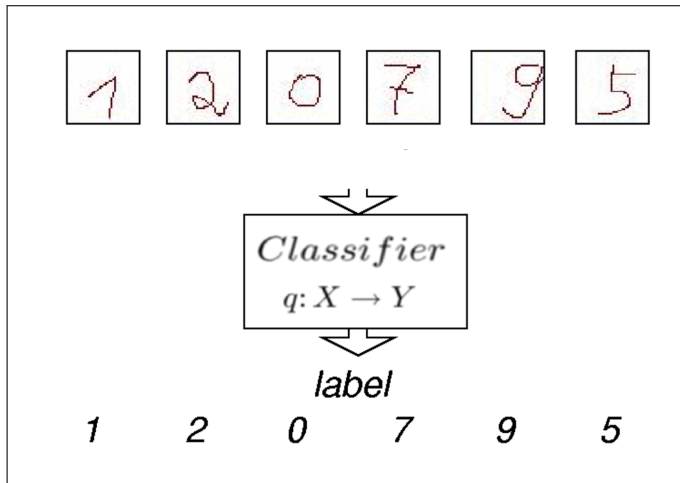


Fig. 1. Illustration of an OCR system [5]

line intersections. Character recognition: The image features will be fed to the neural network and they will be compared with stored glyph features and the nearest match will be chosen, after multiple iterations of classification by the network.

5. APPLICATION

OCR converts images to machine-readable text. That will make it the initial tool that needs to be used for processing any documents or simply any written material in a digital image, which has been captured by a camera[6]. Its output can be stored significantly more compact than scanned images. But beyond that, it enables us to process the output information for numerous applications. Examples of these applications include creating a narrator machine to help the visually impaired read nondigital documents and signs, or automatic recognition of automobile number plates.

6. ACKNOWLEDGEMENT

A very special thanks to Professor Gregor von Laszewski and the teaching assistants Miao Zhang and Dimitar Nikolov for all the support and guidance. This project proposal is written during the spring 2017 semester course I524: Big Data and Open Source Software Projects at Indiana University Bloomington.

REFERENCES

- [1] "Ansible," Web Page. [Online]. Available: <https://www.ansible.com/>
- [2] "What is OCR and OCR Technology," Web Page, 2017. [Online]. Available: <https://www.abbyy.com/en-us/finereader/what-is-ocr/>
- [3] "Playbook," Web Page, Mar. 2017. [Online]. Available: <http://docs.ansible.com/ansible/playbooks.html>
- [4] "A configurable experimental environment for large-scale cloud research," Web Page, Jan. 2017. [Online]. Available: <https://www.chameleoncloud.org/>
- [5] "Demo: Optical character recognition."
- [6] "Optical Character Recognition," Web Page, Mar. 2017. [Online]. Available: https://en.wikipedia.org/wiki/Optical_character_recognition