Optical Character Recognition for Image Files Containing Bisaya Texts

Eirol Jan Coronado
College of Computer Studies
University of the Immaculate Conception
Davao City, Philippines
ecoronado_180000000758@uic.edu.ph

Tristan Montaner
College of Computer Studies
University of the Immaculate Conception
Davao City, Philippines
tmontaner_180000001456@uic.edu.ph

Abstract

Optical character recognition (OCR) is the mechanical or electronic translation of images of hand-written or printed text into machine-editable text [4]. It is performed by optical character readers which are automated electronic systems. OCR may be defined as the process of converting images of machine printed or handwritten numerals, letters, and symbols into a computer- processable format. This study Show the accuracy of the OCR. PyTesseract is the chosen program to assess the accuracy of the Optical character recognition. We used images from different books in order for us to extract text from images. We have also conducted alpha and beta testing to know if we were able to identify if the results will differ if the program was utilized by us or other person. An inconsistent result had been observed while testing the Pytesseract program. Although this program is very easy to use and most efficient, this study is an evidence that OCR is not always 100.

CCS Concepts: • Computing Methodologies → Artificial Intelligence; • Natural Language → Information Extraction.

Keywords: Optical character recognition, text extraction, artificial intelligence, information extraction

Reference Format: Eirol Jan Coronado and Tristan Montaner. . Optical Character Recognition for Image Files Containing Bisaya Texts. In Proceedings of (1st CCS Research Symposium). ACM, Davao City, Philippines, 4 pages.

1 Introduction

Optical character recognition (OCR) is an approach that converts media file input texts into machine-encoded readable texts. Today, OCR is helping in digitizing the manually written archaic compositions, yet in addition helps in changing over the typewritten archives into advanced structure. This has made the recovery of the necessary data simpler as one doesn't need to go through the heaps of records and documents to look through the necessary data. Associations are fulfilling the necessities of advanced safeguarding of note-worthy information, law archives, instructive ingenuity and so forth [3]. Optical Character Recognition, or OCR, is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data. In this fast-paced world, the bank is one of those institutions that use OCR the most. Document digitization in the banking sector is a great utility. Many banks use OCR technology in achieving better transaction security and risk management. The use of OCR software in banks can also scan many customers' important handwritten guarantee documents like their loan documents and more. Additionally, the incorporation of facial recognition software with OCR is also significantly remarkable because it provides two-layer security at ATMs. We need to study Optical Character Recognition so that we will understand its capabilities as well as its strengths and weaknesses. We also need to recognize its full potential so that we can make use of it and so that it will help us make our lives easier. The following twenty years from 1980 till 2000, the product arrangement of OCR was created and sent in instructive organizations, enumeration OCR and for acknowledgment of stepped characters on metallic bar. In the mid 2000s, binarization procedures were acquainted with safeguard authentic

archives in computerized frame and furnish specialists with admittance to these reports. A portion of the difficulties of binarization of chronicled reports were the utilization of nonstandard textual styles, printing clamor and dividing. These applications helped these individuals in creating perusing and composing abilities. In the ebb and flow decade, scientists have dealt with various AI approaches which incorporate Support Vector Machine (SVM), Random Forests (RF), k Nearest Neighbor (kNN), Decision Tree (DT), Neural Networks and so on Analysts joined these AI strategies with picture handling methods to build the exactness of the optical character acknowledgment framework [6]. As of late scientists have zeroed in on creating strategies for the digitization of written by hand reports, basically dependent on profound learning approaches. There are many approaches to do OCR like k Nearest Neighbor (kNN) and Decision Tree (DT) . In this research we evaluate what is the best approach or framework to do OCR for bisaya text.

2 Review of Related Literature

Techtarget explained that OCR systems are made up of a combination of hardware and software that is used to convert physical documents into machine-readable text. Hardware, such as an optical scanner or specialized circuit board, are used to copy or read text while software typically handles the advanced processing. Software can also take advantage of artificial intelligence (AI) to implement more advanced methods of intelligent character recognition (ICR), like identifying languages or styles of handwriting. The process of OCR is most commonly used to turn hard copy legal or historic documents into PDFs. Once placed in this soft copy, users can edit, format and search the document as if it was created with a word processor. OCR makes it possible to make changes to the digital text. What can be done with the digital text depends on which reading software you're using.

Common options include: Highlighting words, sentences or paragraphs, Speaking words aloud using text-to-speech, Changing the colors and the size of text, Placing digital "bookmarks" that enable users to move around within the text. According to [2], Applications that can profit by Optical Character Recognition (OCR) are Banking, Legitimate Industry, Different endeavors, home and office robotization. This application will consolidate the improvement of the outcomes with the utilization of table cutoff points revelation techniques and the utilization of text present getting ready systems on recognizing the uproar and to address dreadful saw words. It's precision of this application is straightforwardly subject to the nature of contribution of the record. One approach to change printed material over to advanced material is by utilizing a scanner. The scanner makes a photograph of the written word. This photograph, regularly called a picture, can be shown on a gadget that has a screen. Yet, checking is just the initial step. The photograph all alone will not empower programming projects to feature words or add different choices that can help your little girl with perusing. This is the place where OCR comes in. A large amount of information exists only as physical documents. Books, forms and other paper-based material is stored and accessed manually. Scanning these paper originals or copies makes it possible to use them in the digital world. Where images contain text, applications such as OCR net can convert the image to usable data. It's OCR that enables number plate recognition for security and toll charging. Cars with speed sign recognition are using OCR. Self-drive cars will use more advanced versions of this to identify and comply with street signs. Translation applications use OCR to capture text from physical documents, signs, and images. The data is then fed into translation software to provide a user with a translation to their own language. Searching large physical documents takes a lot of time and effort. If they're stored as physical documents or photographic images that require extracting from stores, viewing by subject matter experts and possibly specialized equipment to store and view. For example, newspapers are a historical record which can provide detailed information about events. Scanning them with OCR technology makes a search as simple as searching a word file. Once a physical document is scanned using OCR technology the information it holds becomes easily accessible [1]. Documents can provide a resource for creating future documents. If they only exist in a physical form or as images, creating a similar document means starting from scratch. Scanning such a document using OCR technology allows a user to access the text within the document [5]. This can be edited and repurposed with less work than creating a new original document. Physical documents take up physical space. Files, libraries, and storerooms take up expensive space. The storage facility may need to have specialized equipment in order to preserve the documents and protect them from damage or deterioration. Scanned and searchable documents can be stored on conventional data storage devices.

Back up facilities and data security can be managed as it is for other data storage. OCR technology enables artificial intelligence technology to access the huge amount of information that exists in the form of text. This is all around us and forms part of the world we need AI to understand to serve us better. That's why it's so important.

3 Methodology

3.1 Data Collection

This study will focus on the Optical Character Recognition for Image Files Containing Bisaya Texts. The results of this evaluation will be used to support the effectiveness of the system to extract the document files for the aforementioned text. We used PyTesseract in Python on Colab as our method of extraction.

3.2 Materials and Methods

In the data collection, the first thing we did is Install Pytesseract and tesseract-OCR in Google Colab and then we import libraries so that it can recognize the files that you uploaded in the software. Subsequently we Upload the image that we want to extract to the Colab. After that we began Text Extraction to extract the text from the image we uploaded. Finally, we will use the print(extractedInformation) as this will print the text from the image uploaded. The evaluator needs to upload the image that will be converted into text. You need to provide document files for Bisaya text and the image should be comprehensible.

3.3 Data Analysis

The data we have collected from the resources that we have gathered were used to design an application which can help extract bisaya text. We have also considered additional information from the books we have gathered. We acquired the book in the DepEd website. We extracted 10 books, there where 5 books for grade 1, 2 books for kindergarten and 3 books without grade level. In book 1, we extracted 11 pages, in book 2 we extracted 8 pages in book 3 we extracted 7 pages, in book 4 we extracted 14 pages, in book 5 we extracted 10 pages, book 6 we extracted 6 pages, in book 7 we extracted 8 pages, in book 8 we extracted 10 pages, in book 9 we extracted 8 pages, in book 10 we extracted 7 pages.

3.4 Development

We used Google Colab to create a program that would help us extract data from an imported image. In executing the said program, installing necessary libraries and packages to the colab was done. In creating the program, we installed Pytesseract and tesseract-OCR wherein it is a wrapper for Google's Tesseract-OCR Engine as it can read all image types supported by the Pillow and Leoptonica imaging libraries, such as PNG and JPEG. Also, necessary libraries were imported to help the program to properly execute. Then, an image was uploaded to the colab which you can either upload it manually or use the code for uploading the image to he colab. In our case we used the code to upload. With the use of the function image to string it will take the image as an argument and return the extracted text from the image. We used Google Colab to create a program based on the tesseract OCR from a source on the internet to extract data from the imported image or document such as Bisaya texts. The idea is to create a program to extract bisaya text from the imported images.

3.5 Testing

This was the test conducted by the researchers that the prototype system meets the goal objective and lastly the purpose of the study. The test was conducted on the google colab that the researchers test a several images that contains 10 books to evaluate the prototype system. Terms of extracting text from the provided books. In terms of extracting texts from the images acquired from the books, it appears it was not able to extract the text perfectly and it produced a lot of errors such as random text and symbols, although it was able to extract the text perfectly in one out of ten books that we have tried to extract. During the test we have several images that were extracted. We tried that text having an object on it also not printed text. Some of the reasons for the errors are that the text was blurred or written text. Also, it should be a computerized text to extract exactly.

4 Result and Discussion

Book	Pages	Correctly Extracted	Percentage	Remarks
1	11	10	90	The software extracted letter q instead of letter a.
2	8	3	40	The software extracted symbol like the underscore and the equals sign, it also mispelled a word.
3	7	2	30	The software underscore symbol.
4	14	13	90	The software was not able to extract text from a colorful image.
5	10	9	90	The software was not able to extract the numbers
6	6	6	100	All of the pages were correctly extract
7	8	8	100	All of the pages were correctly extract
8	10	10	100	All of the pages were correctly extract
9	8	8	100	All of the pages were correctly extract
10	7	7	100	All of the pages were correctly extract

Table 1. Result and Discussion

5 Conclusion

Optical Character Recognition, or OCR, is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data. Even though it was developed decades ago, it continues to be changed, edited and improved. We extracted 10 books to assess its accuracy. During the extraction, some of the pages were correctly extracted but some were not especially those who have smaller fonts and colorful images. We were able to identify its weaknesses and strengths such as optimizing the process in scanning the image by increasing the scanning resolution so it will be able to extract text from images with letters that have small fonts and to extract text from a colorful image. It also needs to optimize its language dictionary because it's still able to misspell some words. Pytesseract OCR is a very remarkable technology that holds a lot of potential. In this day and age, such software is already quite advanced. However, Pytesseract OCR is going to look even better in the future.

Acknowledgments

First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings throughout my research work to complete the research successfully. We would like to express my deep and sincere gratitude to our research supervisor, Ms. Kristine Mae Adlaon for giving us the opportunity to do research and providing invaluable guidance throughout this research. Her dynamism, vision, sincerity and motivation have deeply inspired us. She has taught us the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honor to work and study under her guidance. We are extremely grateful for what she has offered us. We would also like to thank her for her friendship, empathy, and great sense of humor. We are extremely grateful to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. We are very much thankful to my classmates and friends for understanding, prayers and continuing support to complete this research work. Also we want to express our thanks to our sisters, brother for their support and valuable prayers. Our Special thanks goes to our friends and classmates Jeno Nicoals and Jan Villaflores for the keen interest shown to help us complete this research successfully

References

[1] Faiz Alotaibi, M. Abdullah, R. Abdullah, R. Rahmat, I. A. T. Hashem, and A. K. Sangaiah. 2018. Optical Character Recognition for Quranic Image Similarity Matching. https://www.semanticscholar.org/paper/Optical-Character-Recognition-for-Quranic-Image-Alotaibi-Abdullah/3f0a208882421343eb05b805f8f4cbbc0dbcdcd9

[2] Sumedha B. Hallale and G. Salunke. 2013. Twelve Directional Feature Extraction for Handwritten English Character Recognition. https://www.semanticscholar.org/paper/Twelve-Directional-Feature-Extraction-for-English-Hallale-Salunke/e9edaebf690559ca4dd12c4054e3d341f273f609

[3] A. Mesleh, Ahmed A. M. Sharadqh, Jamil Al-Azzeh, MazenAbu-Zaher, Nawal Al-Zabin, T. Jaber, and Aroob Odeh. 2012. An Optical Character Recognition. https://www.semanticscholar.org/paper/An-Optical-Character-Recognition-Mesleh-Sharadqh/d34c57f398665f3f30d9590f43fb5f09312206bd

[4] Ravina Mithe, Supriya Indalkar, and N. Divekar. 2013. Op-tical Character Recognition. https://www.semanticscholarorg/paper/Optical-Character-Recognition-Mithe-Indalkar/0dc7c304b56d0795064c6b21584450960005a938

[5] Saeeda Naz, A. I. Umar, Saad Ahmed, Riaz Ahmad, S. H. Shirazi, M. I. Razzak, and Amir Zaman. 2018. Statistical features extraction for character recognition using recurrent neural network. https://www.semanticscholar.org/paper/Statistical-features-extraction-for-character-using-Naz-Umar/47b8043d7d713e51bc74bf49edc92cd8dc4e390b

[6] Ömer AYDIN. 2021. Classification of Documents Extracted from Images with Optical Character Recognition Methods. Computer Science, 6(2), 46-55.

CCS
Research Symposium

June 2, 2021

8:30 AM - 5:00 PM

Registration Link:
https://bit.ly/3i40Uvht