

Automated Extraction of Large Scale Scanned Document Images using Google Vision OCR in Apache Hadoop Environment

Rifiana Arief, Achmad Benny Mutiara, Tubagus Maulana Kusuma, Hustinawaty
Information Technology
Gunadarma University
Jakarta, Indonesia

Abstract—This Digitalization of documents is now being done in all fields to reduce paper usage. The availability of modern technology in the form of scanners and cameras supports the growth of multimedia data, especially documents stored in the form of image files. Searching a particular text in a large-scale scanned document images is a difficult task if the document is in the form of images where the text has not been extracted. In this research, text extraction method of large-scale scanned document images using Google Vision OCR on the Hadoop architecture is proposed. The object of research is student thesis documents, which includes the cover page, the approval page, and abstract. All documents are stored in the university's digital library. Extraction process begins with preparing the input folder that contains image documents (in JPEG format) in HDFS Apache Hadoop and followed by reading the image document. The image document is then extracted using Google Vision OCR in order to obtain text document (in TXT format) and the result is saved to output folder in Hadoop Distributed File System (HDFS). The same process is repeated for the entire documents in the folder. Test results have shown that the proposed methods were able to extract all test documents successfully. The recognition process achieved 100% accuracy and the extraction time is twice as fast as manual extraction. Google Vision OCR also shows better extraction performance compared to other OCR tools. The proposed automated extraction systems can recognize text in a large-scale image document accurately and can be operated in a real-time environment.

Keywords—Automation; extraction; google vision OCR; hadoop; scanned document images

I. INTRODUCTION

Document digitization provides an effective way to process, maintain and transfer all types of information from printed form to digital form. The advancement of current information technology and the increased volume of printed documents in many applications, making digitalization of documents increasingly important to reduce paper-based physical documents. This is motivated by the emergence of several issues in the management of physical storage in the form of the risk of damage or loss of paper-based documents and the increasing pile of paper documents that require large storage.

Users in various institutions, such as government, education, medical, commerce and entertainment as well as private companies, have retained documents in electronic form

and at the same time require a fast access service to the desired information[8]. The biggest challenges of large-scale digital document growth are scalability, data consistency, data completeness, time, and security (Chen and Zhang, 2014). Various algorithm has been developed continuously in order to capture, store, search, share, analyze, and visualize data, as well as to anticipate the increased of data volume by increasing the capacity using parallel processing (Chen and Zhang, 2014). Searching a particular text in a large-scale scanned document images is a difficult task if the document is in the form of images where the text has not been extracted.

Text extraction from images can be defined as the work of extracting text objects from a set of images. The results of text extraction can also be used as image search keywords, document search, content-based image search, video content analysis, text-based video search, location search words on documents and others [1]. Text extraction is a challenging task because there are variations of text size, font, style, orientation and alignment to a complex background. Text extraction process from scanned document images includes pre-processing, detection, localization, extraction, enrichment and text recognition. The image to be extracted can be a gray or colored image, in a compressed / uncompressed format. Text detection aims to find the presence of text in images and localization of text aims to find the location of the text and to create boundary boxes of text. The text is then extracted by separating the text from the background image and enriched to improve the quality of the extracted text in order to be recognized. Following the extraction process, the extracted text is recognized using Optical Character Recognition (OCR) [2].

OCR is a technology for recognizing text from images automatically. OCR supports various types of image formats such as JPG, PNG, BMP, GIF, TIFF and PDF files. OCR involves analyzing the captured or scanned images and then translating the image into an editable text format. The text contained in the scanned document image can be easily extracted with the help of an OCR tools. Various OCR applications are available and can be used to extract text on images. Many OCR tools available include Online OCR, Free Online OCR, OCR Convert, Convert image to text.net, Free OCR, i2OCR, Free OCR to Word Convert, Google Docs [6]. The reliability of Google Vision OCR has been shown to extract and recognize text from document images very well compared to other OCRs [5]. The extraction process of the

excessive number of data is automatically performed in Hadoop environment using big data architecture.

Big Data is defined as high volume, high velocity, and/or high variety data sources, which requires new process paradigm to explore the information attached to it, to develop decision-making process, and optimization process. Based on this definition, Big Data is not characterized by specific size metrics, but completion in processing such data because the character (size, velocity or variety) is difficult with conventional processing approach. The Big Data potential is underlined in its definition; but the realization of such potential depends on developing traditional methods or developing new methods capable of handling such large data [4]. Hadoop is a platform for dealing with Big Data and provides problem solving with the ability to analyze large data. Hadoop is an open source based tool that enables distributed processing of large data with multiple clusters to accommodate services. Hadoop is designed to handle data from one service (server) to thousands of machines with high fault tolerance. Parallelization is used for the cost efficiency and processing time required. Big Data includes large-scale, diverse and complex data requiring new architectures, techniques, algorithms and analysis to manage the data and extract the hidden values and knowledge from the data set [3].

In 2013, Tae Ho Hong et. al. developed image-based or pdf-based ebook conversion system to facilitate the search for a words contained in the ebook [7]. To recognize text characters in image files using Tesseract OCR and this conversion process includes large data and uses Map Reduce Hadoop with cluster system so that the conversion process can be successfully done as well as minimizing the processing time. In this research, automated text extraction from large-scale scanned document images based on Google Vision OCR is proposed. The source of documents is stored in multiple folders with different file size using big data technology that is based on Apache Hadoop. The performance measurement of the automated extraction process will be based on the accuracy and the speed of extraction. Extracted results are stored in HDFS to be further analyzed for other purposes.

II. PROPOSED METHODS

Object data used in this research are scanned document images of student's thesis stored in the Gunadarma University library, which includes cover page, approval page, and abstract page as illustrated in Fig.1. The number of documents used in this research are 182,532 with data size of 33.4 GB. All documents are in the JPEG image formats.

The extraction stage aims to extract, to recognize and to get the text contained in the document images. The process starts with reading the input documents using Google Vision OCR and producing the output in the form of text documents. The output text document is stored in HDFS. The extraction process is shown in Fig. 2. The example of automated extraction process is demonstrated in Fig 3.



Fig. 1. Image Document Text in Dataset, (Source: Gunadarma University Digital Library).

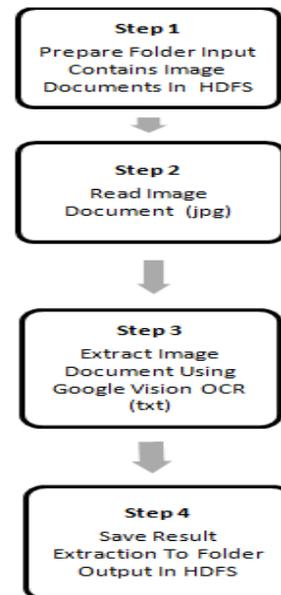


Fig. 2. Method in Extraction Phase.

The automated extraction process for detecting and recognizing text contained in the dataset documents using Google Vision OCR with single node Hadoop is presented in Fig.4. Prior to the extraction, the preprocessing must be done, which includes the preparation of input folder (document images) and to ensure that the document image folder contains document images ready to be extracted. Then, a folder must be created to stored extraction result (document text). All files contained in the input folder is read to get the entire filename, which will be included in the file list. The first document image file is extracted using Google Vision OCR and the fully extracted results is obtained. The extraction results needed are only in the form of text so that objects other than text will be removed or discarded and just take the text as a description then the contents of the description will be saved according to the file name of the scanned document, ie. cov1.jpg becomes cov1.jpg.txt. Following the above process, the results are stored in the HDFS output folder. The extraction process continues to the next document image until the entire contents of the completed folder is processed.

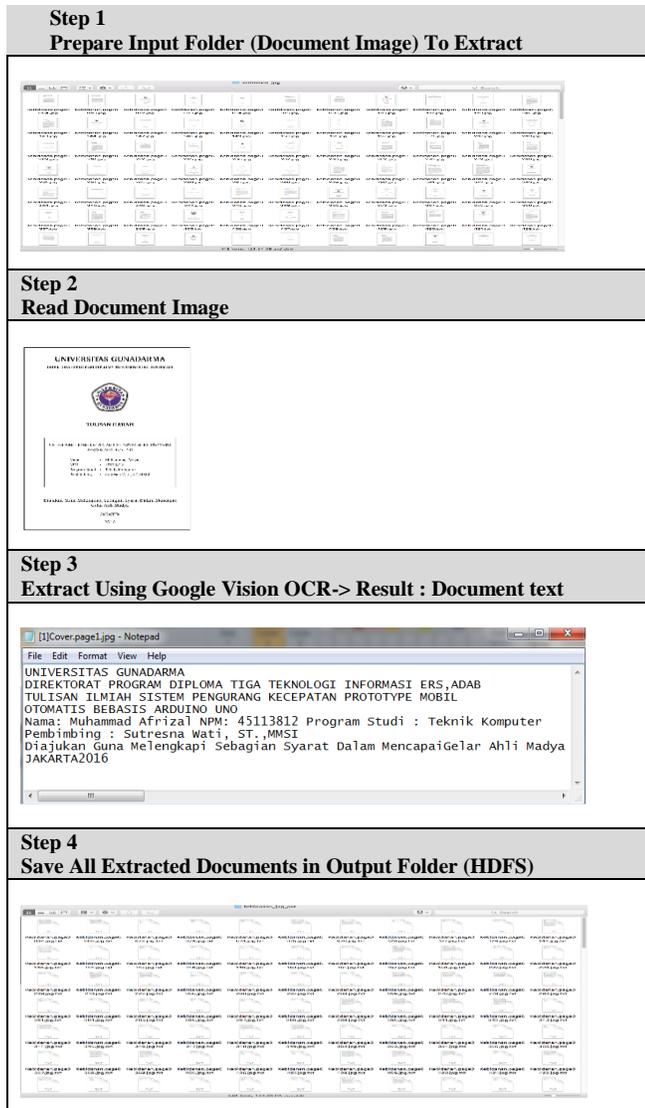


Fig. 3. Example of Automated Extraction Process.

III. RESULT AND DISCUSSION

Performance of the extraction process using Google Vision OCR were tested on three types of document images, namely cover document, approval document and abstract document. The details of the results are shown in Fig. 5, Fig. 6 and Fig.7.

All of the test results from the extraction process are summarized in Table 1. As shown in Table 1, the results of the extraction test using Google Vision OCR on cover documents, approval documents and abstracts have demonstrated successful recognition of the text contents. Limitation on the extraction of cover page has been found during the extraction process where text character in the logo image could not be extracted properly. Therefore logo image was excluded in the process. Limitations on the extraction of approval page has been identified where text character overwritten by signature image could still be read but sometimes could not be accurately recognized. Although there are some limitations found, the

overall performance of the extraction process using Google Vision OCR has shown good results.

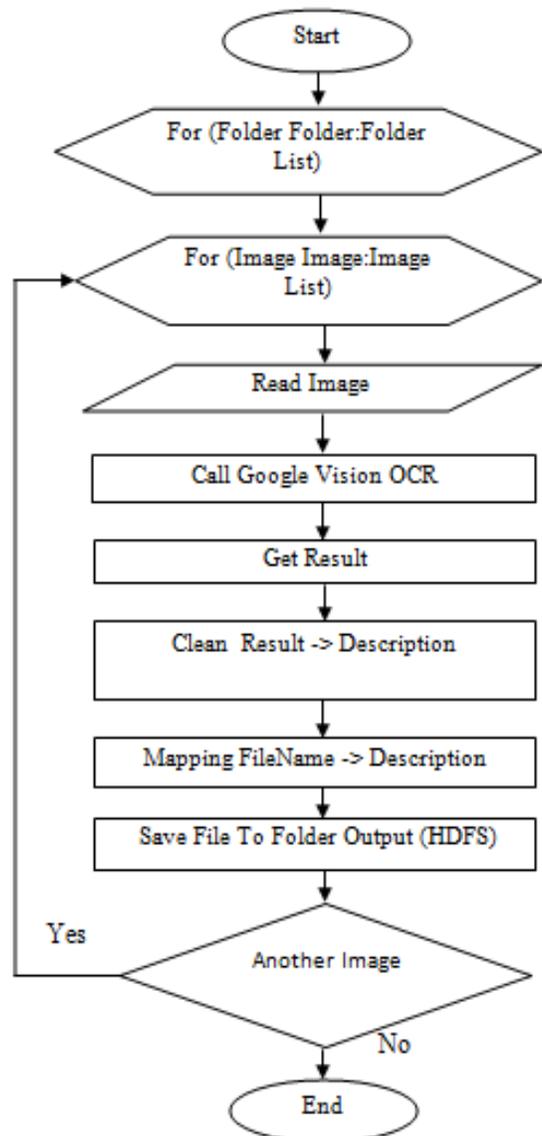


Fig. 4. Extraction Process using Google Vision OCR in Hadoop and Extraction Result Save in HDFS.

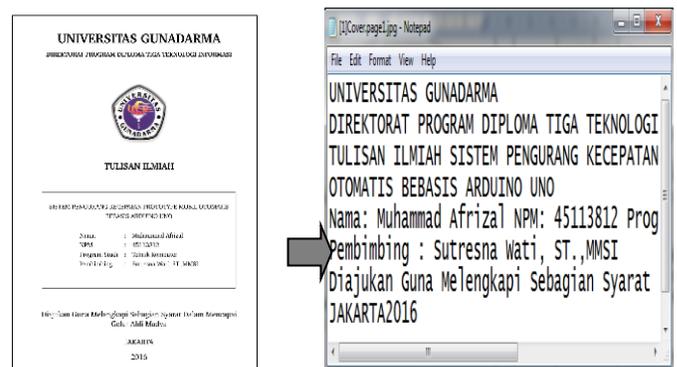


Fig. 5. The Extraction Result of Cover Document Image.

Total extraction time for the entire document (Cover, Approval and Abstract), which is 182,532 files is 15,215 minutes if it was done manually and 7,301 minutes if it was done automatically. In comparing the performance of Google Vision OCR while running manually and automatically, it is shown that the manual process took about 5 seconds to extract while the automated extraction process took about 2.4 seconds to extract a single document. The details are presented in Table 4.

TABLE IV. TIME OF EXTRACTION PROCESS FOR MANUAL AND AUTOMATIC USING GOOGLE VISION OCR FOR ALL DOCUMENTS ON SINGLE COMPUTER

Document Image (jpg)	Extracted Document (txt)	Number of Document	Time to Extract Manual (Minutes)	Time to Extract Automatic (Minutes)
		64,249	5,354	2,570
		59,932	4,999	2,397
		58,351	4,862	2,334
TOTAL		182,532	15,215	7,301

IV. CONCLUSION

This paper presents the implementation of Google Vision OCR to recognize text content in a document image and introduces an automated extraction framework to a large-scale text document image collection in Hadoop architecture. The Google Vision OCR was selected, because it has proven an excellent accuracy compared to other tools. Based on the results of the automated extraction test, the average automated extraction process in Hadoop environment using single

computer is approximately 2 times faster than manual extraction time. All documents in the input folder were successfully extracted, while at the same time the text recognition reached almost 100% accuracy.

Along with the growth of significant data, the future work will be to build an automated extraction system by implementing the automated extraction process using multiple computers in parallel so as to reduce the time required as well as the workload of the computer if using only a single computer. The used of larger datasets with different characteristics will also be considered to see the performance of the proposed system in handling various types of documents. Other types of documents such as government or private data source agencies or documents from the Internet, where scanned document images can be extracted with OCR might also be included. Real-time document retrieval and classification based on text content might also be considered for future work.

ACKNOWLEDGMENT

We would like to thank Gunadarma University Library for their permission to use the students's thesis data.

REFERENCES

- [1] C.P. Sumathi, T. Santhanam and G.Gayathri Devi. "A Survey On Various Approaches Of Text Extraction In Images". International Journal of Computer Science & Engineering Survey (IJCSSES) Vol.3, No.4, August. 2012. Pp 27- 42. 2012
- [2] Divya gera and Neelu Jain. "Comparison of Text Extraction Techniques- A Review". International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2015. Pp 621-626. 2015
- [3] Harshawardhan S. Bhosale, Devendra P. Gadekar. 2014. " A Review Paper on Big Data and Hadoop". International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 pp 1-7
- [4] Beyer, M.A. and Laney, D. "The Importance of "Big Data: A Definition". In Gartner. 2012. <https://www.gartner.com/doc/2057415>.
- [5] H. Choudhary, "Comparing the Top Computer Vision APIs for OCR, in Article Computer Vision, 2017
- [6] S.Vijayarani and A.Sakila. "Performance Comparison Of OCR Tools" . International Journal of UbiComp (IJU), Vol.6, No.3, pp.19-30, 2015.
- [7] Tae Ho Hong, Chang Ho Yun, Jong Won Park, Hak Geon Lee, Hae Sun Jung and Yong Woo Lee, "Big Data Processing with MapReduce for E-Book". International Journal of Multimedia and Ubiquitous Engineering Vol. 8, No. 1, January, 2013 pp. 151-162.
- [8] Yoganand. C.S, Praveen.N, Saranya.N, Ganesh Karthikeyan V., "Survey on Document Classification based on Keyword and Key Phrase Extraction using Various Algorithms". International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 2, February – 2014. Pp 1804-1808. 2014.