Mobile Application for Recognizing Text in Degraded Document Images Using Optical Character Recognition with Adaptive Document Image Binarization

Angie M. Ceniza, Tom Kalvin B. Archival, and Kate V. Bongo
DCIS, University of San Carlos Cebu City, Philippines
Email: amceniza@usc.edu.ph, {archivaltoms, k8bongo}@gmail.com

***Abstract*—Books and documents go through degradation overtime and post threats in the readability of the printed text. Degradations like stains can overlap with the text covering it or ink fading can cause the removal of the text altogether. Converting these texts into digital format can help preserve them. Optical Character Recognition (OCR) is used to transform them into digital text. And, with the increasing computing capability and digital imaging of today's smartphones. We can use them as a convenient tool to capture images of these document and do OCR directly. In this paper, we propose a mobile application that can recognize text in degraded document images using Tesseract as the OCR engine with Adaptive Document Image Binarization to improve the performance of the OCR engine in degraded documents images. The experimental results showed an average character accuracy of 93.17% and word accuracy of 85.82% across 8 degraded document images.**

***Index Terms*—OCR, binarization, degradation, mobile application**

I. Introduction

Documents in libraries or archives present many defects due to aging, ink fading, holes, spots and bleed-through ink [1]. Not only documents, but book and other printed materials also go through that process of degradation overtime. The need to preserve them especially the historical books and documents since they have scholarly and historical information within them. Digital archiving of ancient and historical documents is an expanding trend in heritage study and preservation [2].The transformation of such documents into digital form is essential for maintaining the quality of the originals while provide scholars with full access to that information [3].

In order to transform these documents into digital form, we need to extract the text from these images and make them into a digital text. Optical character recognition (OCR) is an image processing technique that can specifically perform this task [4]. Optical character recognition is also used to enable keyword searches, document categorization, and other referencing tasks [5]. Which provides an advantage to easier archiving and a much better way to search through old documents and books. Given the good quality of the built-in camera and the enough computational capability of smartphones on the market, smartphones can function as portable scanners to extract the texts from the images [6].

The ability of today's smartphone to act as a portable scanner to take images of document and books provide a cost effective and quicker way in digital archiving. The accuracy of standard machine printed OCR system decreases when degradations exist in document images [7]. Preprocessing enhances images for future processing and correct degradations in images [8]. Preprocessing technique's like binarization and segmentation improves the accuracy of OCR [9].

On this study, the researchers aims to develop a mobile application that can do Optical Character Recognition in English or Filipino text in degraded document images. Tesseract will be used as the OCR engine which will handle the character recognition. Adaptive Document Image Binarization will be used as a preprocessing method to improve the accuracy of the OCR engine.

II. Related Literature

A. Optical Character Recognition

Optical Character Recognition is the recognition of printed or written text and converting them into usable machine-encoded text. A study by Singh [10] concentrated on searching for a probable means to cultivate an OCR system for English language when noise occur in the signal and found that neural network gives more accurate result compared to other techniques . Choudhary, Rishi & Ahlawat [11] proposed an approach to extract handwritten character using features pulled out from binarization techniques with a character accuracy of 85.62%. Springman and Ludeling [12] introduced a study on Optical character recognition for Latin text in historical paintings where the OCR engines Tesseract and OCRopus was trained using historical Latin spelling variants that produced an above 90% accuracy in character recognition for 16th century materials.

Manuscript received February 12, 2018; revised June 20, 2018.

doi: 10.18178/joig.6.1.44-47

B. Preprocessing

1) Binarization

Binarization is the conversion of a document image into a binary image which separate the text as the foreground and the background [13]. A study by Sulem, Darbon and Smith [1] compared two image restoration approaches Non-local Means filtering and Total Variation Minimization for the pre-processing of printed documents. Total Variation Minimization is much more effective when the character's degradation is higher. Hedjam, Moghaddam and Cheriet [2] presented an image binarization method using a statistical adaptive approach that uses machine learning classification, a priori information and the spatial relationship of an image. The proposed method preserves weak connections and provides smooth and continuous strokes in degraded document images.

2) Segmentation

Segmentation, used for text-based images, is the process of dividing a digital image into multiple segments (sets of pixels). Its purposed is to retrieve specific information from the entire image where the information could be from a word, line or a character [14]. A study by Gatos and Louloudis [15] analyzed segmenting ancient handwritten document images into text lines and text zones. Text zones were determined by examining vertical lines by rules, white run pixels were used for segmentation and bounding box coordinates for each related component are used to compute the height of a certain character. Choudhary, Rishi & Ahlawat [16] proposed a method to segment character images from text containing images and good results are achieved which shows the strength of the proposed character detection and extraction technique.

3) Adaptive document image binarization

Sauvola and Pietikainen [17] presented a new algorithm for adaptive document image binarization which was tested with document images with degradations. The algorithm calculates a local threshold value using local mean and local standard deviation for each pixel separately. The algorithm is a modified form of Niblack binarization method which give more performance when images background contains light texture, big variations, stained and badly and unevenly illuminated documents [18]. The threshold value is calculated using the formula below.

$$T = m * (1 — k * 1 — S\ R) \quad (1)$$

where T is the local threshold value, m is the mean of pixels under window area, S is the dynamic range of variance and the value of k parameter may be in the range of 0-1. The typical suggested value for k = 0.5 and R = 128.

C. Degraded Documents

The ability of an OCR system to recognize text heavily depends on the printing quality of the input document. A study by Shukla and Banka [19] found that touching characters, broken characters, faxed and heavy prints were the common types of degradation found in Bangla scripts. Droettboom [20] introduced an algorithm for dealing with broken characters in degraded historical printed texts. The algorithm provided good result in rejoining broken characters for the purpose of OCR. Dutta, Sankaran and Jawahar [21] presented a novel character recognition method that resulted in a 15% drop in word error rate on degraded Indian document images. By means of character n-gram images which groups consecutives character segments. Yosef *et al.* [22] proposed variations of methods for accurate recognition and segmentation of highly degraded characters. It improves the known problem on segmentation of historical documents. Namane and Meyrueis [23] presented an OCR method that accomplishes lower errors and was suitable for rejection. The results demonstrate the capability of the model to yield recognition with no errors on poor quality characters even when patterns are highly degraded.

III. Methodology

A. Optical Character Recognition

Tesseract will be used as the OCR engine for the character recognition functionality of the mobile application. It will be implemented using Tess-two, an android library integrated with Tesseract in conjunction with Android Studio for the development of the mobile application.

B. Preprocessing

In order to improve the accuracy and performance of our OCR engine. A binarization method will be use to preprocess the image before it will be pass on to the OCR engine for character recognition. We will be using Sauvola Adaptive Document Image Binarization implemented in the Leptonica Library as our binarization method.

C. Extracting Text

- Tesseract starts with line finding using a line finding algorithm which involves a page layout analysis that identify the text area and marks them with a baseline.
- Once the baseline is identified, the next step is to identify the words and chops them into characters using fixed pitch detection.
- The characters are then classified using a Static Character Classifier and match the characters with its actual character.
- The new words are then form and is check if it's a true word using Linguistic Analysis.
- The sentences are then reform with the new words.

D. Testing and Validation

The researchers opted to measure the character accuracy and word accuracy to determine the effectiveness of the OCR mobile application in recognizing text in degraded documents [24] with eight document images with different types of degradation. The

character accuracy will be measured using the formula below.

$$(\#characters - \#errors) / \#characters \quad (2)$$

While the word accuracy will be measured using the formula below.

$$(\#word - \#errors) / \#word \quad (3)$$

where the #character is the total amount of characters, #word is the total amount of words, #error for both character and word is the Levenshtein distance which is the minimum number of edits (insertions, deletions or substitutions) to transforms wrong characters or words to their correct form.

IV. Result and Discussion

The following document images in Fig. 1 are the samples used to test the word and character accuracy.

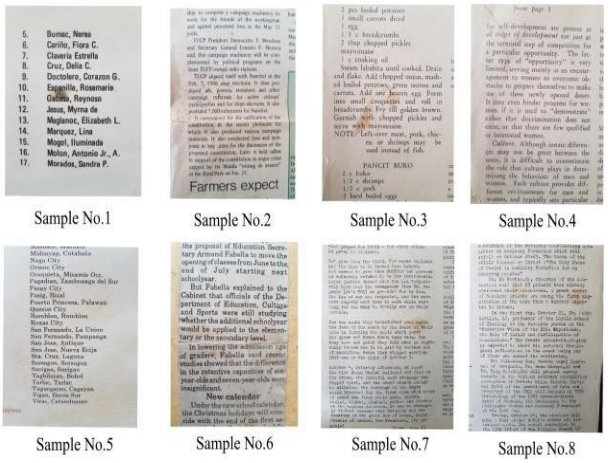

Figure 1. Sample document images used.

Table I and Table II shows the character accuracy and word accuracy of the mobile application in extracting text on the sample document images. Where the sample no. indicates as to what image it is, langauge is the langague that the text are written in and the degradation type define's what degradation is mostly present in the image.

TABLE I. Character Accuracy

Input			Output
Sample No.	Langauge	Degradation Type	Character Accuracy
1	Filipino	Background Stains	95.70%
2	English	Background Noise	86.53%
3	English	Background Stains	94.00%
4	English	Background Stains	90.61%
5	Filipino	Background Stains	98.29%
6	English	Background Stains & Misaligned text	90.07%
7	English	Background Stains & Heavy Print	95.94%
8	English	Background Stains & Broken Character	94.27%

TABLE II. Word Accuracy

Input			Output
Sample No.	Langauge	Degradation Type	Word Accuracy
1	Filipino	Background Stains	91.17%
2	English	Background Noise	82.05%
3	English	Background Stains	83.33%
4	English	Background Stains	80.72%
5	Filipino	Background Stains	96.55%
6	English	Background Stains & Misaligned text	81.25%
7	English	Background Stains & Heavy Print	81.50%
8	English	Background Stains & Broken Character	90.00%

Table III shows the comparison of the proposed mobile application compared to two other methods. The proposed mobile application produced a higher average character and word accuracy compared to the other two.

TABLE III. Comparison with Other Method's

Method	Average Character Accuracy	Average Word Accuracy
Tesseract w/ Sauvola	93.17%	85.82%
Tesseract w/ Otsu	74.39%	76.95%
Abbyy OCR Engine	91.45%	76.34%

V. Conclusion and Future Works

The study used Tesseract as its OCR engine and Sauvola Adaptive Document Image Binarization for the preprocessing and aimed to create a mobile application that could do Optical Character Recognition in degraded document images. Based from the result, the mobile application performance is acceptable being able to recognizing text in degraded document images with an average character accuracy of 93.17% and word accuracy of 85.82%. The study can be further improved by using a different OCR engine since it seems that Tesseract word accuracy seems to decrease as more broken characters are present in the document image. Another improvement is the OCR of handwritten text since the current study doesn't include it.

Acknowledgment

We would like to thank the Department of Computer and Information Sciences, School of Arts and Sciences and University of San Carlos for the research opportunity.

References

[1] L. Likforman-Sulem, J. Darbon, and E. H. B. Smith, "Pre-processing of degraded printed documents by non-local means and total variation," in *Proc. 10th International Conference on Document Analysis and Recognition*, Barcelona, 2009, pp. 758-762.

[2] R. Hedjam, R. F. Moghaddam, and M. Cheriet, "A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images," *Pattern Recognition*, vol. 44, no. 9, pp. 2184-2196, 2012.

[3] E. Kavallieratou and E. Stamatatos, "Improving the quality of degraded document images," in *Proc. 2nd International Conference on Document Image Analysis for Libraries*, Lyon, 2006, p. 10 & p. 349.

[4] S. Sharma and R. Sharma, "Character recognition using image processing," *IJAETMAS*, pp. 115-122, 2006.
[5] M. Gupta, N. Jacobson, and E. Garcia, "OCR binarization and image pre-processing for searching historical documents," *Pattern Recognition*, pp. 389-397, 2007.
[6] X. Du, W. Fan, J. Wang, Z. Peng, and M. Sharaf, "Web technologies and applications," in *Proc. 13th Asia-Pacific Web Conference*, 2011.
[7] M. K. Jindal, R. K. Sharma, and G. S. Lehal, "A study of different kinds of degradation in printed Gurmukhi script," in *Proc. International Conference on Computing: Theory and Applications*, Kolkata, 2007, pp. 538-544.
[8] M. Sonka, V. Hlavac, and R. Boyle, "Image pre-processing," in *Image Processing, Analysis and Machine Vision*, Boston, MA: Springer, 1993.
[9] J. Liang, C. Patel, A. Patel, D. Patel, A. A. Shinde, and H. Wu, "Optical character recognition by open source OCR tool tesseract: A case study," *International Journal of Computer Applications*, vol. 55, no. 10, 2012.
[10] S. Singh, "Optical character recognition techniques: A survey," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 2, no. 6, 2013.
[11] A. Choudhary, R. Rishi, and S. Ahlawat, "Off-line handwritten character recognition using features extracted from binarization technique," *AASRI Procedia*, vol. 4, pp. 306-312, 2013.
[12] U. Springmann and L. Anke, *OCR of historical Printings with An Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus*, 2014.
[13] Y. Chen and L. Wang, "Broken and degraded document images binarization," *Neurocomputing*, vol. 237, pp. 272-280, 2017.
[14] B. Ankita and K. Mali, "A comparative study on image segmentation based on artificial bee colony optimization and FCM," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 3, pp. 1284-1394, 2014.
[15] B. Gatos and G. Louloudis, "Segmentation of historical handwritten documents into text zones and text lines," *Proc. ICFHR*, vol. 9, p. 464, 2014.
[16] A. Choudhary, R. Rishi, and S. Ahlawat, "A new approach to detect and extract characters from off-line printed images and text," *Procedia Computer Science*, vol. 17, p. 434440, 2013.
[17] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225-236, 2000.
[18] P. More and D. Dighe, "A review on document image binarization technique for degraded document images," *International Research Journal of Engineering and Technology*, vol. 3, no. 3, 2016.
[19] M. K. Shukla and H. Banka, "A study of different kinds of degradation in printed Bangla script," in *Proc. 1st International Conference on Recent Advances in Information Technology*, Dhanbad, 2012, pp. 119-123.
[20] M. Droettboom, *Correcting Broken Characters in the Recognition of Historical Printed Documents*, pp. 364-366, 2003.
[21] S. Dutta, N. Sankaran, and P. K. C. V. J. Robust, "Recognition of degraded documents using character N-grams," *IEEE*, 2012.
[22] I. Bar-Yosef, A. Mokeichev, K. Kedem, I. Dinstein, and U. Ehrlich, "Adaptive shape prior for recognition and variational segmentation of degraded historical characters," *Pattern Recognition*, vol. 42, no. 12, pp. 3348-3354, 2009.
[23] A. Namane and P. Meyrueis, "Multiple classifier for degraded machine printed character recognition. Antoine Tabbone et Thierry Paquet," *Colloque International Francophone sur l'Ecrit et le Document*, France, Groupe de Rechercheen Communication Ecrite, pp. 187-192.
[24] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fourth annual test of OCR accuracy. [Online]. Available: http://www.expervision.com/wpcontent/uploads/2012/12/1995 The_Fourth_Annual_ Test_of_OCR_Accuracy.pdf

Angie M. Ceniza is an assistant professor in the University of San Carlos currently serving as the Computer Science Coordinator of the Department of Computer and Information Sciences. Being a recipient of CHED Scholarship grant, she obtained her degrees of master in information technology in University of San Jose Recolitos (October 2012). She obtained her doctor of philosophy in technology management in Cebu Technological University (November 2017). She has been an author of think, click & share: a comprehensive worktext in media and information literacy for senior high school. She has been a peer reviewer in local and international conferences. And, has published her research in referred and scopus-indexed journal. Her research interest includes data mining, natural language processing, image processing and artificial neural networks.

Kate V. Bongo is currently studying computer science at the University of San Carlos. She was born on November 11, 1997. She lives in Nasipit Talamban, Cebu City. She used to be a working scholar last 2014 - 2017. In her work, she used to be an assistant at the registrar's office last 2014-2015 and was transferred as a Laboraty Assistant at the Department of Computer and Information Science. And also, took her OJT last 2017 - 2018 as a Web Develoer at Konserygo. And now, she's an SK (Sangguniang Kabataan) Councilor of Barangay Talamban.

Tom Kalvin B. Archival was born on February 4, 1997 in Cebu, Philippines. He is a student taking his bachelor degree in computer science at the University of San Carlos. His research interest includes natural language processing and image processing.