Verifying the Integrity of Hardcopy Document Using OCR

S. Mthethwa
Women in science without borders
smthethwa@csir.co.za

N. Dlamini
Women in science without borders
ndlamini2@csir.co.za

Abstract

Hardcopy document forgery is still a challenge and occurs frequently nowadays. Many countries have reported a lot of cases, including South Africa where government issues documents are forged. Protecting these documents from being tampered with is necessary at all times. Various methods have been presented to deal with the challenge of document forgery such as, e.g. Optical Character Recognition (OCR). In this paper, we improve OCR with the aim to achieve a high accuracy to eliminate the misrepresentation of characters read from an image file. To implement the solution we use an OCR tool, Tesseract. The experimental setup is explained and the results which yielded an accuracy of 100% are discussed in detail. While this is on-going work, the experimental results demonstrate the feasibility of using OCR as part of the solution.
.

Themes— Digital Revolution.

1 Introduction

The use of Optical Character Recognition (OCR) in document analysis has dispersed widely as a field of research and a tool to recognize characters. This is due to its quality to convert and recognise text that is present in a digital image file and the continuous improvements aimed at achieving an impeccable standard of text recognition accuracy [1]. Optical mechanisms enable the computer application utilized to recognize characters, part of this process also requires pre-processing of the image files and training the system used to accumulate knowledge of the characters in the selected language, e.g. English. The gained knowledge is then used by the trained system to recognise the text [2].

Researchers continue to present many uses for OCR e.g. licence plate recognition, document analysis, etc. To compensate for the inconsistencies in OCR such as, not achieving a 100% accuracy rate when recognizing characters, they use other methods such as hash functions, digital signatures, 2D Barcodes, etc. Hard-copy document authentication and verification is a subject that is widely active in research, because despite the rapid growth of paperless environments, certain matters still need to be communed in paper document form. For instance, birth certificates, passports, insurance documents, driving licenses, or case files [3], but these files can be forged or faked. Husain A. et al. [4] proposed a system that can verify if a hard-copy document has been tampered with, in this system they include the use of OCR as a component to check the integrity of the document contents (e.g. text) but also highlight that human inspection is also required due to the characters that are misrepresented. In a case where OCR has failed and required human intervention, this paper aims to improve character recognition by enhancing the knowledge gained by Tesseract, an OCR tool in order to achieve a 100% recognition accuracy to allow an automatic document verification process.

2 Method

To conduct the experiments, a tool was developed that could utilize OCR for character recognition. The tool is divided into two parts, namely; document definition and validation, which are discussed below.

2.1 Document Definition

This is a process whereby, a document is defined in order to create a dataset. These documents are saved as XML-based files, referred to as meta-template. The meta-template consists of coordinates (height and width) of the document to be defined. Each text in the meta-template is labelled with a unique identifier, which makes it easier for the process of validation. The meta-template consist of two types of text; normal text and validation text (important text that must be verified). This template is then used to generate a matching-template (which is to be used during the process of validation), only consisting of validation text.

2.2 Document Validation

This is a process whereby, a copy of the document is checked against the original document to determine whether they are the same or it has been tampered with. To validate the pdf documents, the tool reads in an image (which can be obtained either by, converting the pdf documents to images or by printing and scanning the pdf documents and saving them as images). To validate, the tool reads in the image and pass it through to OCR which

returns a string of characters. The output string (from OCR) is compared to the expected string (from the matching template) and the results are returned to the tool. Therefore determining the integrity of the document.

3 Results

To conduct the experiments, a dataset containing 100 documents was generated using AnyOCR font (optimized for OCR applications which only consists of capital letters and numbers). AnyOCR font is described as the best OCR font [5]. Results are depicted in Table 1 and 2 below.

Table 1: AnyOCR Experimental Results

Type	Total No. of Images	Accuracy (%)
Whole document	100	100%
Text labels	3500	100%
Characters	38 362	100%

Table 2: AnyOCR Scanned Results

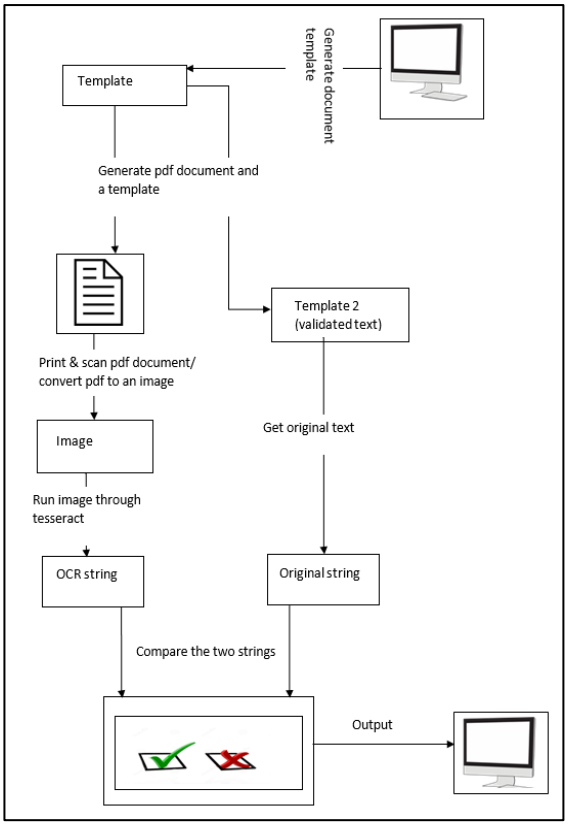
Category	Accuracy (%)	Error Rate (%)
Whole Document	96.0%	4.0%
Text Labels	100%	0%
Characters	99.997%	0.003%

Our tool managed to achieve an accuracy of 100% for original documents, although we did not manage to achieve that for immediate printed and scanned documents. As this is an on-going work, we still need to improve our tool for accurate results and incorporate other fonts as well as working with damaged documents.

4 Conclusion

This paper presented our proposed solution using OCR for verifying the integrity of a hardcopy document. Documents were generated using a font known as AnyOCR and Tesseract was used to validate the documents. The experimental results yielded an accuracy of 100%, which is good. Since this is still ongoing work, Tesseract is still going to be trained on multiple fonts as well as damaged documents, with the expectation of high accuracy results and the ability to verify text automatically without human intervention.

A Appendix



Acknowledgments

The authors would like to express their gratitude to the CSIR Modelling and Digital Science for sponsoring this research.

References

[1] G. Vamvakas, B. Gatos, and S. J. Perantonis, "Handwritten character recognition through two-stage foreground sub-sampling," *Pattern Recognit.*, vol. 43, no. 8, pp. 2807–2816, 2010.

[2] S. Singh, "Optical Character Recognition Techniques : A Survey," *J. Emerg. Trends Comput. Inf. Sci.*, vol. 4, no. 6, pp. 545–550, 2013.

[3] M. Warasart and P. Kuacharoen, "Paper-based Document Authentication using Digital Signature and QR Code," in *2012 4TH International Conference on Computer Engineering and Technology (ICCET 2012)*, 2012.

[4] A. Husain, M. Bakhtiari, and A. Zainal, "Printed document integrity verification using barcode," *J. Teknol. (Sciences Eng.*, vol. 70, no. 1, pp. 99–106, 2014.

[5] M. Jenckel, S. S. Bukhari, and A. Dengel, "AnyOCR: A sequence learning based OCR system for unlabeled historical documents," in *Proceedings - International Conference on Pattern Recognition*, 2017, pp. 4035–4040.