IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 3, No. 3, September 2014, pp. 121~128
ISSN: 2252-8938

Optical Character Recognition of Off-Line Typed and Handwritten English Text Using Morphological and Template Matching Techniques

Olakanmi Oladayo O
Department of Electrical & Electronic Engineering, University of Ibadan Nigeria

Article Info

Article history:

Received May 23, 2014
Revised Aug 9, 2014
Accepted Aug 21, 2014

Keyword:

Handwritten English Text
Morphological
Off-Line Typed
Optical Character
Recognition
Template Matching Techniques

ABSTRACT

The existence of several documents in historical archives which need to be edited and stored in a computer has been one of the drives of Optical Character Reader (OCR) research. Earlier scanner has been used to achieve this tedious task however scanner only produces picture images of the documents.This makes the documents unreadable and un-editable through other word processing applications.This paper proposed an OCR system which converts off line typed and handwritten texts into their editable textual representations.The morphological correlation technique improves the mapping and recognition efficiency of the OCR system.

Corresponding Author:

Okalanmi Oladayo O,
Departement of Electrical and Electronic Engineering,
Rm. 6, Faculty of Technology,
University of Ibadan Nigeria.
Email: olakanmi.oladayo@ui.edu.ng

1. INTRODUCTION

It has become a trend to document most of the documents in the archives using scanner, however, these documents cannot be edited or read thereafter by computer systems. Due to the fact that scanner scans documents as an image not as encoded set of characters. OCR system does electronic translation of handwritten or printed text into machine encoded text. OCR is widely used to convert books and documents into electronic files and to computerize a record keeping system in an office. OCR makes it possible to edit such document, search for a word or phrase, store it more compactly, display or print a copy and apply techniques such as machine translation, text-to-speech and text mining to it. OCR study was started by Tyurin a Russian scientist (A.Jain and Karu 2006). The first modern character recognizers appeared in the middle of the 1940s with the development of the digital computer. The early work on the automatic recognition of characters has been concentrated either upon well printed text or upon small set of well distinguished handwritten text or symbols, although, successful but had been implemented mostly for Latin characters and numerals. Also some studies on Japanese, Chinese, Hebrew, Indian and Arabic charades and numerals in both printed and handwritten cases were also considered by some OCR systems. The developments in OCR until 1980s suffered from lack of advanced algorithm, powerful computing hardware and optical devices. With the outward explosion on the computing technology development, the previously proposed methodologies found a fertile environment for rapid growth in many application areas. Presently, renewed vigours are being put in the optical character recognition research. One of these is recognition of printed and handwritten documents. More sophisticated algorithms which utilize advanced methodologies are

being developed. In this work two methodologies are combined to achieve an efficient OCR system which will be able to recognize off-line typed and handwritten documents.

The remaining part of this paper is arranged as follows: section 2 is the review of related works on OCR systems and methodologies. The design methodology and working principle of the propose OCR system are explained in section 3. Section 4 contains the test results and conclusion.

2. RELATED WORK

OCR can be described as one of the applications of pattern recognition. It wide acceptability is due to various existing documentation challenges which OCR systems are able to solve. OCR systems can be classified according to two classification metrics; data acquisition method and text/ language type. The OCR methodologies depend greatly on the type of the equipment used for data acquisition and the kind of text the data are represented with. OCR data acquisition systems may be online or off-line data acquisition system. Off-line data acquisition captures data from paper through optical scanners or cameras whereas on-line data acquisition systems use the digitizer which directly capture writing with the order of the strokes, speed, pen up and down information. On-line OCR is adaptive in the sense that immediate feedback is given by the writer whose corrections can be used to further train the recognizer. Apart from this, it involves very little processing. Operation such as smoothing, segmentation, de-slanting, de-skewing and feature extraction operations such as line orientation, loops corners and cusp detection are easier with the pen trajectory data than on pixel images. However, on-line OCR system requires a special pen and tabloid which are not comfortable and natural to used as pen and paper. Apart from this, it cannot be used to convert printed or handwritten documents on papers (Ullmann 1987).

Off-line OCR system does recognition on the bits pattern for both printed and handwritten text. The bit pattern is represented by a matrix of pixels. This matrix may be of large size. In order to make the pattern consistent most of the scanners are standardized to 100 to1600 dots per inch (Fukunaga 1990). Most of the research works are on off-line OCR systems because it allows previously printed or handwritten texts to be processed and recognized. Some of the developed off-line OCR systems are postal address reading, cheque sorting, short hand transcription, reading aid for visual-impaired.

Various research works had been done on variety of methodologies that are used in OCR systems. Not only this, several works had been done on various applications of OCR such as plate number recognition, different languages text recognition. For example, (Mohammed n.d.) used template matching approach to identify Musnad characters which is considered as basis for Arabic language. He extracted and normalized Musnad characters from input image. The extracted character was compared to each template in the database to find the closest representation of the input character using 2-D correlation coefficient approach.

In (Huang, Learned-Miller and McCallum n.d.) cryptogram algorithm was engaged to implement OCR system. Cryptogram algorithm groups similar characters in the document and solves a cryptogram to assign labels to clusters of characters. With this method, no character model is needed and can arbitrarily handle any font styles. However, it was discovered that this approach cannot handle numerals, punctuation marks and uppercase.

In (Kamaljit and Balpreet May 2013), a morphological approach was adopted to identify plate number. Their implementation was able to identify the first character of the plate number.

3. OCR METHODOLOGY

OCR is the science that entails the description or classification of character measurements that usually based on some models. OCR is one of the categories of image recognition. There are various character recognition methods used in developing character recognizer. These methods are: neural network, moment based approach, contour based approach, template matching and morphological approach. In this work template matching and morphological techniques are used to recognize English texts.

Template matching refers to the process of detecting an object having a certain size, shape and orientation in an image by applying an operator containing positive weights in a region resembling the objects to be detected and containing negative weights in a region surrounding the positive weight (R.M.K Sinha 1997). Morphology as derived from biology is a branch of biology which deals with the form and animals and plants. It is adopted in this context as a tool for extracting image components that are useful in the representation and description of the region shape. There are several procedural steps engaged in achieving morphological techniques. These include filtering, thinning, pruning, erosion and dilation, opening and closing.

3.1. TEMPLATE MATCHING AND MORPHOLOGICAL TECHNIQUE

Template matching and morphological techniques as stated earlier, are OCR recognition techniques. These algorithms involve features extraction and classifier. In template matching image pixels are used as the features being extracted from both the input character and the classified characters. The classifier compares the input character features with a set of character template in the character class. In this context the character class contains numerals, upper and lower cases of English characters as shown in fig 1 and fig. 2. The absolute value of the classifier procedure which is the correlation coefficient between the input character and the considered character template is used to morphologically determine the template with a closest correlation match.

Formally,

$$X = (U, L, N, P) \quad (1)$$

$$c_n = \{y \in (U, L)\} \quad (2)$$

where:
U is the set of uppercase English characters
L is the set of lowercase English characters
N is the set of English numbers
P is the set of English punctuation marks
The transformation function δ on character c is:

$\tau_n : c_n \; X \; \delta \; \rightarrow \; \tau_n$

$\tau \; is \; the \; set \; of \; \; templates \; of \; characters$

English characters are classified into numerals, upper and lower cases. In the character class some of the characters were written in different ways in order to accommodate different ways of writing. This OCR system, as shown in figure 3, is grouped into three processing levels which are low level processing, intermediate level and high level processing. These are implemented using 64-bit Matlab version 7.8.0.387 and the input texts are built with paint brush and text.

3.1.1. LOW LEVEL PROCESSING

As shown in the figure 3, low level processing involves image acquisition and pre-processing of the acquired images. Image acquisition stage acquires image of the document or characters to be recognized. Most time input character image is of finite resolution which ultimately affects the quality of its transformation, therefore, pre-processing becomes necessary. The pre-processing stage includes colour normalization, scaling filtering and thinning. Colour normalization is used to change input character foreground colour to black and background colour to white. To achieve this, histogram technique was used. The input character was used to form histogram of single class which was grouped into intervals. Over each of these intervals a vertical rectangle is drawn with its area proportional to the number of point falling into that interval. The luminance of the image was determined using equation 3. Figure 2 and 3 depict the input image before and after normalization respectively.

$$Lu = 0.3R + 0.59G + 0.11B \quad (3)$$

Normalization algorithm:

1. Select the relevant part of the character.
2. Determine the threshold for the colour normalization
3. Process the image from top corner line by line
4. Store the R,G,B value of each pixel
5. Determine Lu using equation 1
6. If Lu< threshold value then turn the pixel black otherwise white.
7. Repeat for the whole input image

The image scaling scales the input character image up or down depending on the original size. This was done to reduce the recognition time and error rate as large character images would take longer time to process while small image may be difficult to recognize. After scaling the character becomes blocky and hence the smoothening filtering stage removes the spike edges. This stage contains smoothening filter, low

pass filter. These filters are used to reduce blurring and noise. Also, implemented in the low level processing is the thinning which converts any elongated parts or strips in the image regardless of their bits into narrow strips that are only about one pixel wide.

3.1.2. INTERMEDIATE LEVEL PROCESSING

Intermediate Level Processing (ILP) in the in figure 3 involves image rotation and segmentation. Sometimes input character image may not be properly aligned in angular fashion with respect to the character template set. An instance of this will be corrected by realign the image OCR. Segmentation which forms the core of IL processing stage partitions the input image into its constituent characters. Shown below is the algorithm used for segmentation

Segmentation algorithm:

1. Scan the image from right to left row wise
2. Add and count all the x coordinates
3. Determine the x-coordinate of the centroid using $xcentroid = \sum(x)/n$ where n is the total number of the centroid.
4. Determine the y-coordinate of the centroid using $ycentroid = \sum(y)/n$ where n is the total number of the centroid.

3.1.3. REPRESENTATION AND DESCRIPTION

Representation maps the scanned character image to form suitable for subsequent computer processing while description is a feature selection which deals with extracting features in some quantitative manner or differentiating one class of objects from another. This was achieved using internal characteristics, that is, the pixels compromising the region.

3.1.4. KNOWLEDGE BASE

The knowledge base contains the numbers, punctuation, upper and lower cases of English alphabets as shown in Figure 4a-4b. It is basically a database of typed and handwritten English alphabets, numbers, and punctuations. Individual character images in the knowledge base are used to generate the correlation values for the input character image and output character text.

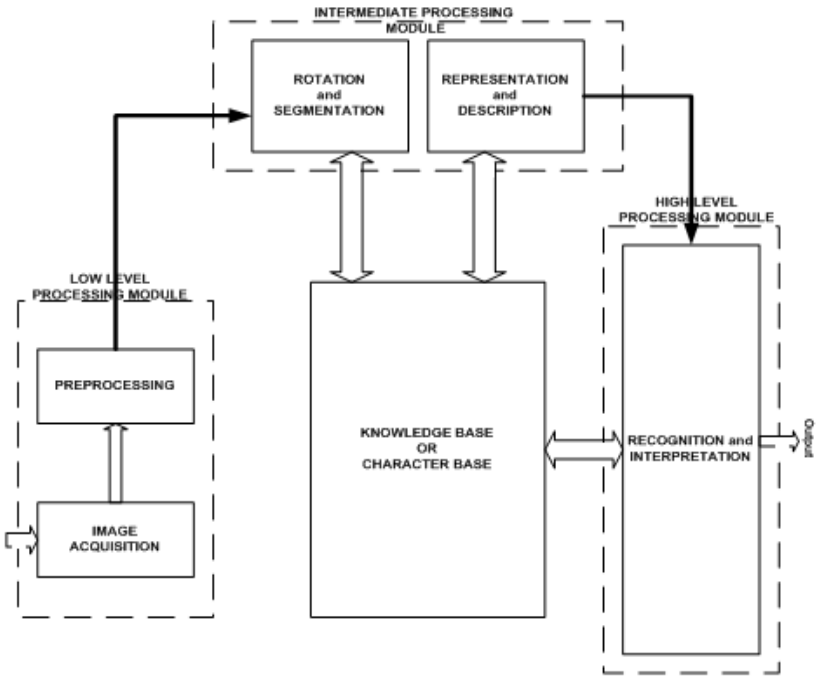

Figure 1. Schematic of the off-line Optical Character Reader

Figure 2a. Input image character before normalization

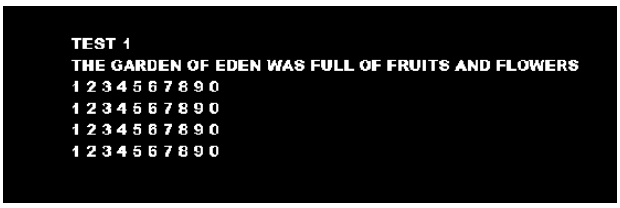

Figure 2b. Input image character after normalization

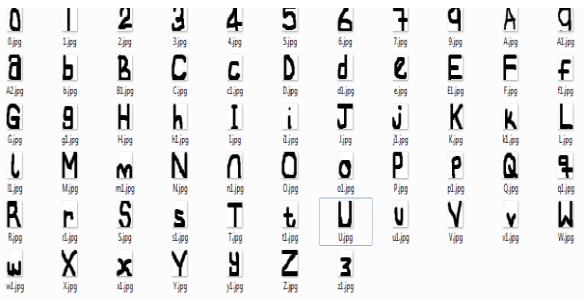

Figure 3a. OCR handwritten English character knowledge base

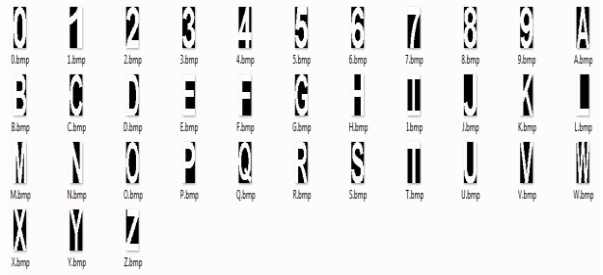

Figure 3b. OCR typed English character knowledge base.

4. TEST AND CONCLUSION

The OCR system was subjected to different set of input text images in order to determine its recognition efficiency. The test was carried out on both typed and handwritten input texts. The input images as shown in Figure 4a, 5a and 6a are different set of input texts created using the paint brush as pen and paint text which represent handwritten and typed English texts respectively. The outputs of the OCR system for the input text image in fig 4a, 5a and 6a are shown in figure 4b, 5b and 6b respectively. The test results were quite impressive. It was observed from the OCR output in fig. 4b that character *G* was the only character not recognized. This shows an accuracy of 99% for the typed text with execution time of 112 char/sec recognition rate. Also, for input text in fig. 5a it was observed from the OCR output in fig. 5b that just a few number of characters were not properly recognized (*I,G, comma, space)*. The OCR system output in fig. 6b which represents OCR output for the handwritten input text in fig. 6a, recorded an accuracy of 90%. It was observed that the OCR system's performance unit is independent and constant for handwritten and typed text

images of different size. Also, the result showed that the developed OCR system more effectively recognized numerals than alphabets.

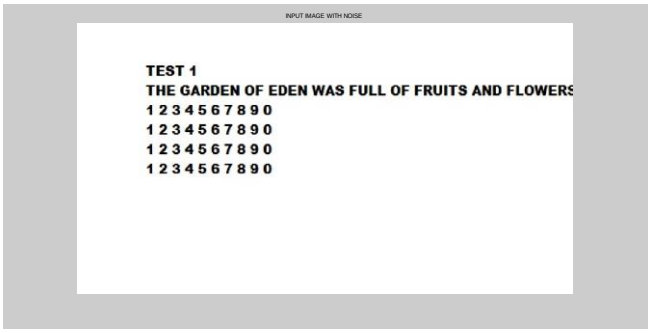

Figure 4a. OCR input of a scanned image text document.

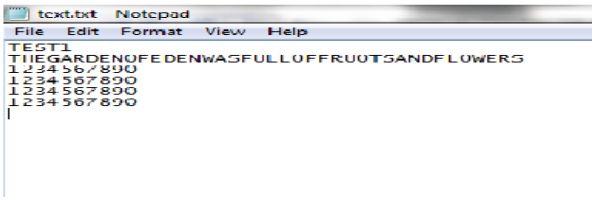

Figure 4b. OCR output of the scanned image text document in 5a.

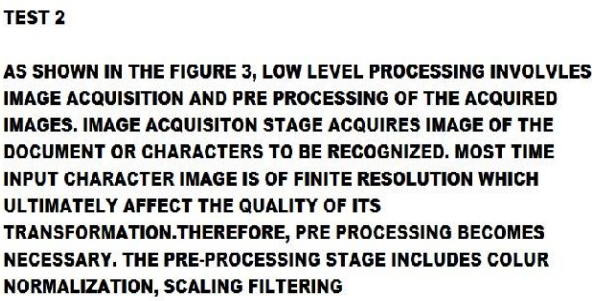

Figure 5a. OCR input of a scanned image text document

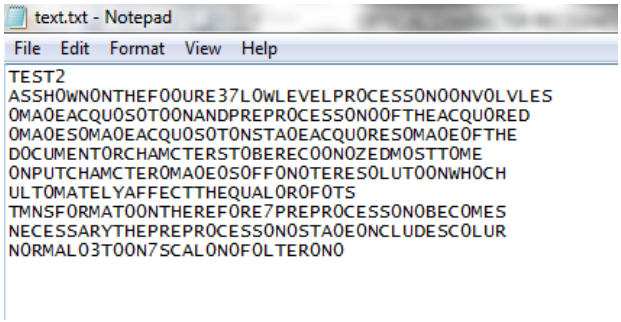

Figure 5b. OCR output of the scanned image text document in 5a

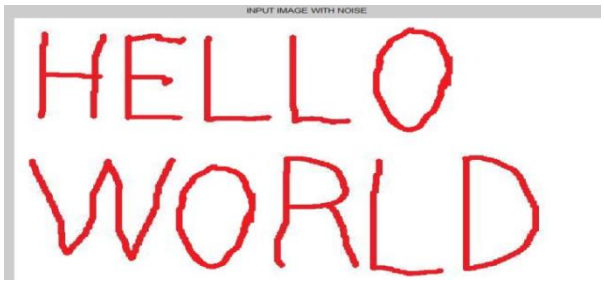

Figure 6a. OCR input of a scanned handwritten image text document.

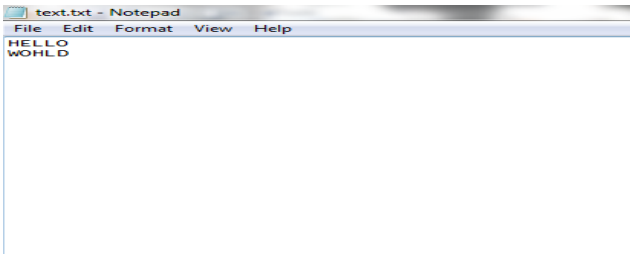

Figure 6b. OCR output of a scanned handwritten image text document.

REFERENCES

[1] A.Jain, and K. Karu. "Page Segmentation Using Texture Analysis, Pattern Recognition." 29 (2006): 743-770
[2] Dueire Lins, R, G Pereira Silva, and A.R Gomes e Silva. *"Assessing and Improving the Quality of Document Images Acquired with Portable Digital Cameras." ICDAR 2007*. Ninth International Conference on (Volume:2). Document Analysis and Recognition, 2007. 569-573
[3] Fukunaga, K. *Introduction to Statistical Pattern Recognition.* 1990
[4] Huang, Gary, Erik Learned-Miller, and Andrew McCallum. "Crytogram Decoding for Optical Character Recognition."
[5] Huang, Kaizhu, Jun Sun, Y. Hotta, and K. Fujimoto. *"An SVM-Based High-accurate Recognition Approach for Handwritten Numerals by Using Difference Features." ICDAR 2007*, Ninth International Conference on Document Analysis and Recognition. 589-593
[6] Kamaljit, Kaur, and Kaur Balpreet. "Character Recognition of High Security Number Plates Using Morphological Operator." *International Journal of Computer Science & Engineering Technology (IJCSET)* 4, no. 5 (May 2013)
[7] Kundu, A., McLean MITRE Corp., T. Hines, J. Phillips, and B.D. Huyck. *"Arabic Handwriting Recognition Using Variable Duration HMM." ICDAR 2007.* Ninth International Conference on. Document Analysis and Recognition, 2007
[8] Lin, Shang-Hung. "An Introduction to Face Recognition Technology." *Informing Science special issue on Multimedia Informing Technologies* Vol. 3, no. 1 (2000)
[9] Mohammed, Ali Q. "Template Matching Method for Recognition Musnad Characters based on Correlation Analysis."
[10] Nadeem, Danish, and Saleha Rizvi. "Character Recognition Using Template Matching." M.sc Project
[11] Nawaz, Tabassam, Syed Ammar Hassan, Shah Naqvi, Habib ur Rehman, and Anoshia Faiz. "Optical Character Recognition System for Urdu (Naskh Font) Using Pattern Matching Technique."
[12] Pratap, R.L., L. Satyaprasad, and A. Sastry. *"Middle Zone Component Extraction and Recognition of Telugu." ICDAR 2007*, Ninth International Conference on Document Image Document Analysis and Recognition
[13] Qing, Chen, and Petriul M Emi. "Optical Character Recognition for Model-based Object Recognition Applications."
[14] R.M.K Sinha, et.al. "HybridContextual Text Recognition with String matching." *Pattern Analysis and Machine Intelligence (PAMI).* 1997. 915-925
[15] Saqib, Rasheed, Naeem Asad, and Ishaq Omer. *"Automated Number Plate Recognition Using Hough Lines and Template Matching."* Proceedings of World Congress Engineering and Computer Science WCECS. 2012
[16] Sun, Jun, Kaizhu Huang, Y. Hotta, and K. Fujimoto. *"Degraded Character Recognition by Complementary Classifiers Combination." ICDAR 2007.* Ninth International Conference on . Document Analysis and Recognition, 2007
[17] Ullmann, J.R. *Application of Pattern Recognition.* CRC Press, Inc., 1987
[18] Yin, Xu-Cheng, Jun Sun, S. Naoi, and K. Fujimoto. "A Multi-Stage Strategy to Perspective Rectification for Mobile Phone Camera-Based Document Images." 2007

BIBLIOGRAPHY OF AUTHORS

O.O Olakanmi received the B.Tech in Computer Engineering from Ladoke Akintola University of Technology, Ogbomosho 2000, M.sc in Computer Science and Ph.D. in Electrical and Electronic Engineering from University of Ibadan. He is a lecturer in the Department of Electrical & Electronic Engineering, University of Ibadan and major in Data Communication, Parallel & Distributed Computing.