Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 11 No. 1, 2020 - www.cisdijournal.net

An Efficient Technique for Text Extraction with Self Organizing Property in Big Data

[1]Bennett, E.O., [2]Elujekor, V. & [3]Taylor, O.E.
Department of Computer Science
Rivers State University
Port Harcourt, Nigeria
E-mails: [1]bennett.okoni@ust.edu.ng , [2]velujekor@gmail.com, [3]onate.taylor@ust.edu.ng

ABSTRACT

Our world is at an inflection point, one defined by Big Data and business analytics. For most, getting value from big data is a challenge as data is being generated every second and stored in relational databases. This paper is aimed at developing an efficient text extraction technique with self-organizing properties in Big Data. The text extraction is achieved through segmentation, feature extraction and classification. Data organizing is accomplished by utilizing Self-Organizing Map algorithm; in view of the standards of vector quantization and proportions of vector similarity. Self-organizing property has been developed for the clustering of input vectors and has been used as unsupervised learned classifiers. However, the Self-Organizing property algorithm further groups the information in blocks, lines and words by giving unified access to data residing in multiple sources. The implemented System eliminated non-textual document, extracted textual data and organized them in blocks, lines and words thus, ensuring that data is minimized and could be used in different forms.

Keywords: Text Extraction, Data Fusion, Optical Character Recognition, Self-Organizing Property

CISDI Journal Reference Format
Bennett, E.O., Elujekor, V. & Taylor, O.E. (2020): An Efficient Technique for Text Extraction with Self Organizing Property in Big Data. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 11 No 1, Pp 1-10.
Available online at www.isteams.net/cisdijournal

1. INTRODUCTION

Many organizations today, are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as “big data” because of its volume, the velocity with which it arrives, and the variety of forms it takes. Big data is creating a new generation of decision support data management. Businesses are recognizing the potential value of this data and are putting the technologies, people, and processes in place to capitalize on the opportunities. Gathering and storing Big data makes little worth; it is just data foundation now. It must be analyzed and the outcomes are utilized by decision makers so as to generate approvals. According to [1], Big data is defined as high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Integrating image text is the process of fusing multiple records representing the same real-world objects into a single, consistent, and clean representation [2]. [3] describes the classification of data fusion techniques and outlined the progress on schema mapping, record linkage and data fusion by data integration community. It addressed the challenges of Big Data Integration and identified a range of open problems for future research community.

Optical Character Recognition belongs to the family of machine recognition techniques and automatic identification. To capture knowledge, an electrical device is used that converts the particular image or sound into a digital file. Text detection and recognition in pictures and videos could be an analysis space that makes an attempt to develop an Automatic Data Processing (ADP) system with the power to mechanically scan from pictures and videos the text content visually embedded in advanced backgrounds [4]. One of the advantages of information integration systems is that the user of such a system obtains a complete yet concise overview of all existing data without needing to access all data sources separately: complete because no object is forgotten in the result; concise because no object is represented twice and the data presented to the user is without contradiction. The latter is difficult because information about entities is stored in more than one source. After major technical problems of connecting different data sources on different machines are solved, the biggest challenge remains: Overcoming semantic heterogeneity; that is overcoming the effect of the same information being stored in different ways.

The main problems are the detection of equivalent schema elements in different sources (schema matching) and the detection of equivalent object descriptions (duplicate detection) in different sources to integrate data into one single and consistent representation. However, the problem of actually integrating, or fusing, the data and coping with the existing data inconsistencies is often ignored. Text Extraction is the process of mining multiple records representing the same real-world object into a single representation. The focus of this research is on text extraction systems with self-organizing properties (Block, Line and Word). This contribution thus proposes an approach, which relies on optical character representation, for the automatic design of text extraction system.

2. RELATED LITERATURE

Text extraction is a wide-ranging subject and many definitions are available in the literature. It is a component of data fusion. [5] stated that data fusion can be defined as any process of aggregating data from multiple sources into a single composite with higher information quality. The research by [6] gives comprehensive survey of different techniques and how the techniques can be scaled up for larger scenarios. It divides the sensor data fusion into three main categories. Complementary technique means putting together information from different sensors to get a bigger picture. Redundancy category fuses data from similar sensors to increase the accuracy of data. Cooperative operations combine to create new information. It describes combining different sensor networks and combines their data to form a predictive contextually rich model. The author also describes the evaluation framework for various data fusion techniques and compares various solutions based on the evaluation framework. The research does not propose any solution based on the findings of the study and does not compare the computational cost and delay introduced in the communication.

[7] developed an optical character recognition system based on image preprocessing technologies combined with Least Square Support Vector Machine (LS-SVM). It uses dynamic threshold operation and robust gray value normalization to segment characters and extract features respectively, and then uses LS-SVM to classify characters based on features. The system was evaluated by carrying out recognition experiments on the optical characters of electronic components. [8] proposed automatic text location in images and video frames, to perform a color reduction by bit dropping and color clustering quantization, and afterwards, a multi-value image decomposition algorithm is applied to decompose the input image into multiple foreground and background images. [9] proposed a system for automatic number plate recognition which detects the vehicle, captures the vehicle image, extracts the vehicle number plate region using image segmentation and recognizes the characters on number plate using Optical Character Recognition (OCR). The recognized registration number is then compared with the database and the vehicle's owner is discovered.
[10] proposed an automatic text extraction system, where second order derivatives of Gaussian filters followed by several non-linear transformations are used for a texture segmentation process. Then, features are computed to form a feature vector for each pixel from the filtered images in order to classify them into text or non-text pixels.

3. SYSTEM DESIGN

System design shows the overall framework and levels representation of a system. The captured image is extracted through segmentation method; extracted symbols are preprocessed, features are extracted, and post processed output text into block, line and words.

3.1 Segmentation

The segmentation stage takes in an image and separates the different parts of an image, like text from graphics, lines of a paragraph, and characters of a word. In Segmentation, optical recognition function first segments the page into paragraphs, which are subsequently split into lines. Picking character; segmentation splits text from graphics. When the text is segmented, it isolates characters or words. The mostly occurred problem in segmentation is that it causes confusion between text and graphics in case of joined and split characters. Usually, splits and joints in the characters are caused by scanning. If document is dark photocopy or if it scanned at low threshold, joints in characters will occur. And splits in characters will occur if document is light photocopy or scanned at high threshold. System also gets confused during segmentation when characters are connected to graphics. Figure 1 shows details of character segmentation process.

0	0	0	0	0
0	1	1	0	0
0	1	1	0	0
0	1	1	0	0
0	0	0	0	0

Figure 1: Matrix Representation of Image

Figure 1 shows character segmentation which uses matrix consisting of 5 rows and 5 columns with a cell representation of 0 and 1. Cell containing letter is 1 and empty cell is 0

$$Percentage\ of\ real\ image = \frac{image\ size}{image\ frame} * 100\% \quad (1)$$

$$Percentage\ of\ real\ image = \frac{6}{25} * 100\%$$ as seen in figure 1.

3.2 Preprocessing

When a camera captures image on paper, some blur image sections may occur during scanning process. This results in poor recognition of characters. This problem is overcome by pre-processing. It consists of smoothing and normalization. In smoothing, certain rules are applied to the contents of image with the help of filling and thinning techniques. Normalization is responsible to handle uniform size and slant of characters. Normalization changes n-dimensional grayscale image. Intensity is the quantity of Red Green Blue (RGB) bits in an image.

For example, if the intensity range of the image is Red (50), Green (80) and Blue (75), the desired range is given as:

R= $\frac{50}{256}$ = 0.1, G = $\frac{80}{256}$ = 0.3, B = $\frac{75}{256}$ = 0.2

4 FEATURE EXTRACTION

The feature extraction stage is used to extract the most relevant information from the text image which helps to recognize the characters in the text. The selection of a stable and representative set of features is the heart of pattern recognition system design. Feature extraction is the process of transforming the feature set to a new feature subspace with lower dimensionality than the original. Using techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), data compression is performed by summarizing the original information into lower dimensions. As the datasets have the correct format, features can be selected for the analysis phase. In doing so, the amount of data to process is decreased, which also decreases the complexity. Selection of feature sets can also help handling common challenges, specifically, the objective of feature selection is three-fold: improving the performance of classifiers, providing faster and more cost-effective classification, and providing a better understanding of the underlying process that generates the data. Blindly selecting features may not yield an optimal subset of features, which then decreases efficiency.

Therefore, methods for feature selection exist, in two dimensional moments of order (a+b) of a gray level of binary image can be defined as:

$$m_{ab} = \sum_{y,x} y^a x^b f(y,x) \qquad (2)$$

where a,b = 0,1,2,.....∞

The function f (y,x) provides pixel value of yth column and xth row of the image.

The sums are taken over all the pixels of the image. The central moments with translation invariance of order (a+b) can be written as:

$$m_{ab} = \sum_{yx} (y - \tilde{y})^a \, (x - \dot{x})^b f(y,x) \qquad (3)$$

Where
$\tilde{y}$ = M10/M00, where 1 represent character and 0 represent no character
$\dot{x}$ = M01/M00

5. POST PROCESSING

After the extraction stage if there is any word which is unrecognized, then the word is given a meaning in this stage. The overall fusion processes are shown in Figure 2.

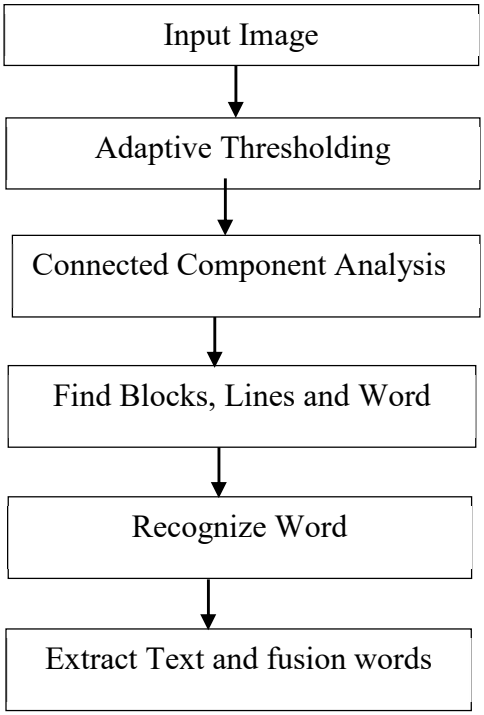

Figure 2: Extraction processes

- **Input image**: The image given as the input is a gray or RGB (Red, Green, Blue) image. The input image must be a flat image that is a parallel image capture. It doesn`t have any capability to rectify the errors caused due to perspective distortion.

- **Adaptive thresholding**: It converts the gray scale image to binary image and calculates the optimal threshold so that there is minimal variance difference between the background and foreground pixels.

- **Connected component analysis**: It searches for the foreground image and treats them as blob. Blob refers to the region in the digital image which differ in comparison to the surrounding due to different colour or brightness.

- **Line finding**: Lines are found by analyzing the image space adjacent to the potential character. If the pixel count is below a specified threshold level, then it is detected as line.

- **Recognize Word**: After all characters have been extracted it recognizes word line by line.

System Setup
To conduct the experiment, the following minimum requirements is expected:

Hardware Requirements:
Processor: Intel Centrino 1.6Ghz Processor or higher or other equivalent processors.
Memory: 1GB of (RAM) or higher ; Camera: 13MP Camera with flash

Software Requirements:
Operating system: Windows 7 and above
Android Studio; Min SDK 19
Target SDK 28, Java programming language, Database: Firebase

6. RESULTS AND DISCUSSION

Five sets of scanned input document were used. Camera scanned object and picked multiple textual documents in blocks as shown in figures 5 to 8. The organized textual documents in lines is shown in figures 9 and 10 while the organized textual document in words property is shown in figure 11. However, Table 1 shows the extraction of multiple data, and gives details of the words extracted and integrated; after its integration in blocks and lines the system combined the data in words and places words in rows to form a correct sentence. Table 2 shows the statistics of scanned documents, it consists of five scanned documents (1,2,3,4,5): scan 1 has the highest number of words and sentences before and after extraction with 00.11 seconds; data integration time is dependent on the number of words and sentences in the document.

1. $\text{Percentage of text extraction} = \frac{133}{583} * 100 = 22.8\%$
2. $\text{Percentage of text extraction} = \frac{113}{509} * 100 = 22.2\%$
3. $\text{Percentage of text extraction} = \frac{83}{890} * 100 = 9.3\%$
4. $\text{Percentage of text extraction} = \frac{48}{421} * 100 = 11.4\%$
5. $\text{Percentage of text extraction} = \frac{36}{518} * 100 = 6.9\%$

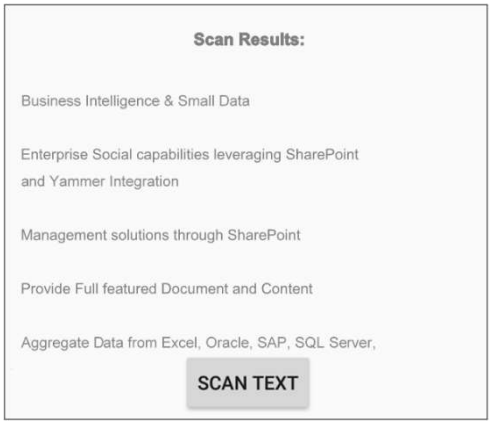

Figure 5: Data Integration in Block (1)

Figure 6: Data Integration in Block (2)

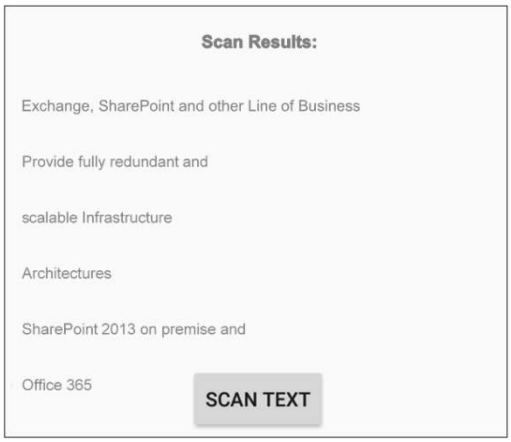

Figure 7: Data Integration in Block (3)

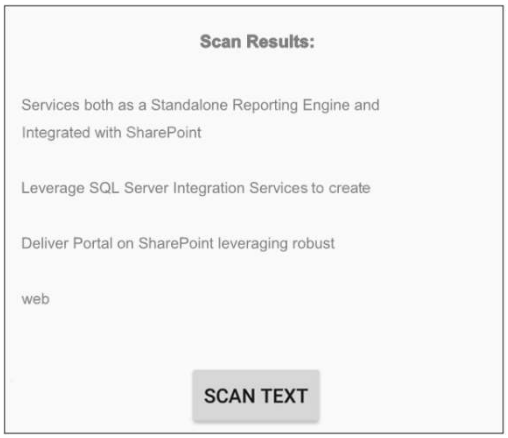

Figure 8: Data Integration in Block (4)

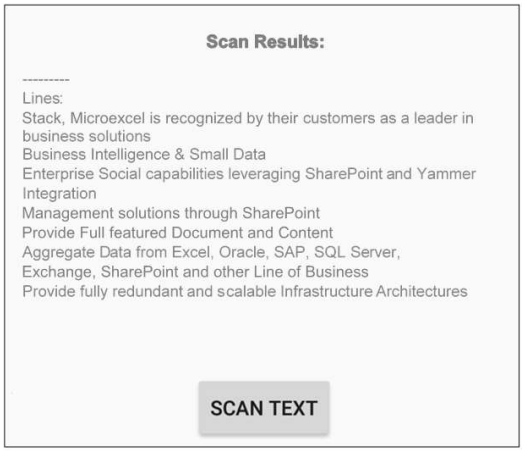

Figure 9: Line (Organizing Property) (1)

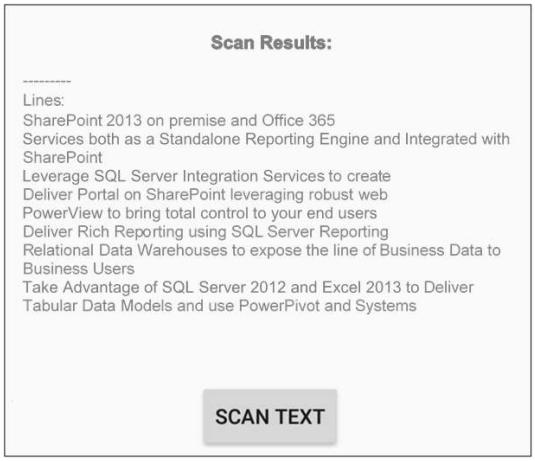

Figure 10: Line (Organizing Property) (2)

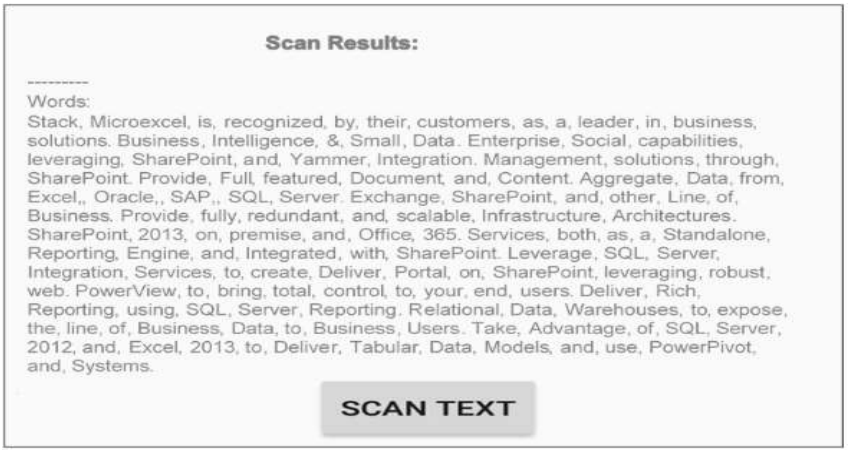

Figure 11: Word (Organizing Property)

Table 1: Extraction of sample multiple documents

S/N	SCANNED LINES
1	Stack, Microexcel is recognized by their customers as a leader in
2	Business solutions
3	Business Intelligence & Small Data
4	Enterprise Social capabilities leveraging SharePoint and Yammer
5	Integration
6	Management solutions through SharePoint
7	Provide Full featured Document and Content
8	Aggregate Data from Excel, Oracle, SAP, SQL Server,
9	Exchange, SharePoint and other Line of Business
10	Provide fully redundant and scalable Infrastructure Architectures
11	SharePoint 2013 on premise and Office 365
12	Services both as a Standalone Reporting Engine and Integrated with
13	SharePoint
14	Leverage SQL Server Integration Services to create
15	Deliver Portal on SharePoint leveraging robust web
16	PowerView to bring total control to your end users
17	Deliver Rich Reporting using SQL Server Reporting
18	Relational Data Warehouses to expose the line of Business Data to
19	Business Users
20	Take Advantage of SQL Server 2012 and Excel 2013 to Deliver Tabular
21	Data Models and use PowerPivot and Systems

Table 1: Extraction of Multiple Data

Table 2: Time Log For Different Scanned Document

NO OF SCAN	NO OF LINES BEFORE EXTRACTION	NO OF LINES AFTER EXTRACTION	WORDS BEFORE EXTRACTION **(B)**	WORDS AFTER EXTRACTION (B)	EXTRACTION DURATION (S)
1	27	21	169	138	00.11
2	20	15	140	118	00.10
3	22	17	129	101	00.9
4	17	11	116	82	00.7
5	25	18	98	73	00.7

Table 3: Data Size Before and After Extraction

NO OF SCANNED	DATA SIZE BEFORE EXTRACTION (KB)	DATA SIZE AFTER EXTRACTION (KB)
1	583	133
2	509	113
3	890	83
4	421	48
5	518	36

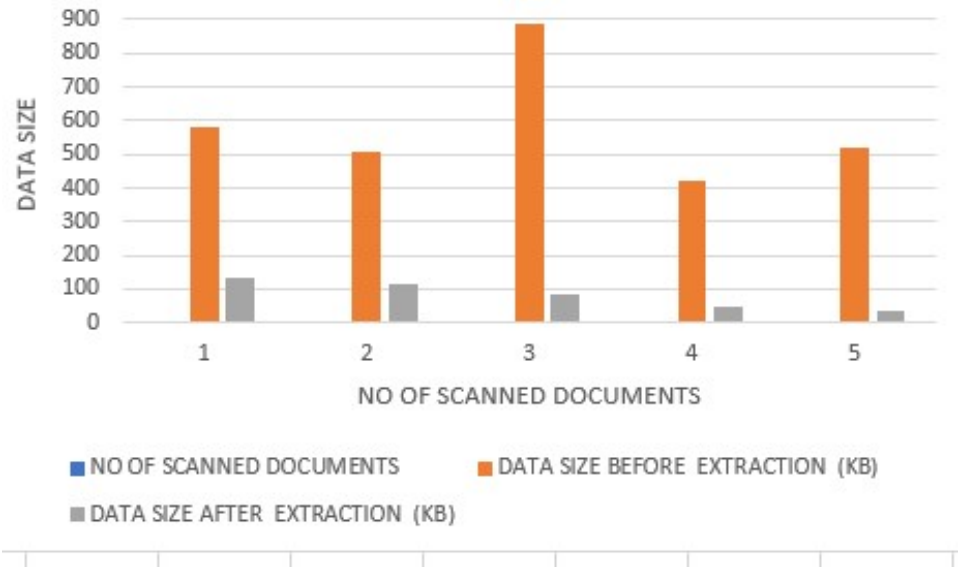

Figure 12: Data Size Before and After Extraction

7. CONCLUSION

The system scanned a document, extracted and organized text into blocks, lines and words using a vision camera. In this paper, we have successfully integrated multiple data in meta data and considered the following: elimination of non-textual data from the original document, thus, final document is a percentage of the original document in terms of size. The document size after integration is lesser than the size before integration, thus, data is minimized and can be used differently.

REFERENCES

[1] Loshin D. (2013). Big Data Analytics, *Elsevier*, pp. 142.
[2] Bleiholder, J. & Naumann, F. (2008). Data fusion. ACM Comput. Surv., 41, 1, Article 1, 41 pages DOI,10.1145/1456650.1456651
[3] Castanedo F. (2013). A Review of Data Fusion Techn Techniques. Deusto Institute of Technology, Volume, Article ID704504,19pages
[4] Dong X. L., & Srivastava D. (2013). "Big Data Integration": Data Engineering (ICDE), *2013 IEEE 29th International Conference*, pp. 1245–1248.
[5] Lehal G.S. and Singh C. (2000). A Gurmukhi Script Recognition System. *Proceedings of the International Conference on Pattern Recognition (ICPR),* 1051-4651/00.
[6] Ribeiro, R. A., Falcão, A., Mora, A., & Fonseca, J. M. (2014). A fuzzy information fusion algorithm based on multi-criteria decision making. *Knowledge-Based Systems, Intelligent Decision Support Making Tools and Techniques: {IDSMT}*.58(0):23 – 32.
[7] Wang, M., Perera, C., Jayaraman, P., Zhang, M., Strazdins, P., Shyamsundar, R., & Ranjan, R. (2016). City Data Fusion: Sensor Data Fusion in Internet of Things. *International Journal of Distributed Systems and Technologies*, 7(1), 15-36.
[8] An Y., Zou Z. & Li R. (2016) Descriptive. Characteristics of Surface Water Quality in Hong Kong by a Self-Organizing Map. *International Journal of Environmental Research and Public Health*, 13(1), 115.
[9] Xie, J. (2009). Optical Character Recognition Based on Least Square Support Vector Machine", *Third International Symposium on Intelligent Information Technology Application (IITA)*, pp. 626-629.
[10] Jain, A. K. &Yu, B. (1998). Automatic Text Location in Images and Video Frames. *In Proceedings of International Conference of Pattern Recognition (ICPR),* pp. 1497-1499.