Document Feature Extraction Based on Unoccupied Space Using Triangle Model: A Preliminary Work

A. Tahir[1,2], M. S. Azmi[2], N. Ahmad[2], N. A. Arbain[2] and A. R. Radzid[2]
[1]Department of Information & Communication Technology,Politeknik Ungku Omar,
Jalan Raja Musa Mahadi,
31400 Ipoh, Perak, MALAYSIA.
[2]Computational Intelligence and Technologies Research Group (CIT-Lab), C-ACT
Faculty of Information & Communication Technology,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 71600,
Durian Tunggal, Melaka, MALAYSIA.
sanusi@utem.edu.my

***Abstract*— Document identification is used to extract information from a digital document such as Al-Quran, articles, agreement and so on. With increasing digital documents on the internet, it is important to identify that the document is genuine or not. There is existing research on document identification. However, the problem occurs when character recognition is done for particular language only and it is hard to recognize the character when the image dataset size was small. Therefore, the purpose of this research is to make use the similarities of each character language which is unoccupied space as the document feature extraction using triangle model. As for the preliminary work, the objective is to obtain a list of point selection that will be used for triangle model from the generated histogram.**

This research using an experimental design, the dataset was chosen is own dataset which document image will be used and IFN/ENIT dataset in order to handle the small size dataset. While the techniques involve is Otsu's Model and histogram normalization. Experiments were conducted on own dataset word segment of documents. The results were able to obtain a list of point selection for both histograms vertical and horizontal. This tool is able to recognize document from other language documents.

***Index Terms*— Document identification; Histogram normalization; Otsu's Model; Unoccupied space.**

I. INTRODUCTION

Nowadays, documents been digitized and are available for access all over the world through internet. However, there are plenty of documents which are unnamed and also been copied by others author. Therefore, there is a need to identify the original document and original author known or unknown.

Document identification is the special case in which one document been analysed whether all parts of the document were written by the same author (known or unknown) [1]. It also a task that works with data digitalization and extraction which the classes be defined to represent particular types of documents such as digital documents [2]. Generally, it is one of the areas that use to distinguish the objects in a document and to check its originality.

Research on document identification keeps growing eventually using existing, improved or new techniques in image processing and the type of recognition such as character recognition [3], handwriting recognition [4], word recognition [5] and etc. Then, the extraction and classification process will be performed to obtain the result.

The dataset in the most research is in a critical issue. This is due to lack of standard database [3] which the researcher needs to develop their own database and small database size [2] because existing developed dataset is a local dataset and only applicable for their own research. This makes it difficult for other researchers to use the same data set as a benchmark in order to verify the algorithm. Hence, [6] told that to get the accuracy of propose algorithm in this research, the nearest standard dataset can be considered.

Second issues are existing research uses mean average of pixel coordinates to form a triangle towards sets of data digit and single form calligraphy [7]. However, this situation cannot be applied to the extraction of the document because the document structure is much different from a single character [6] and [8]. Therefore, does the existence of unoccupied space in the document can be counted as part of feature extraction? If not, then need a new method to produce a feature of the exploitation of unoccupied space.

In this paper as preliminary work, we propose to obtain a list of point selection that will be used for triangle model from histogram generated. For the first step, the document is scan into the image. Next, thresholding techniques which are Otsu's model been selected to convert document image into a binary image consisting of '1' and '0'; 1 for unoccupied space and 0 for occupied space. The binary image will be analyzed and map into histogram vertical and horizontal. Finally, point selection list is obtained from the maximum occurrence for each data from both histograms.

The paper has the following structure: In the second section, presents some related works on document identification, unoccupied space and thresholding. The third section contains a description of preprocessing phase. In the fourth section, the experiments and results of the research will be discussed and in conclusion, a summary of results and some plan for future work is written.

II. RELATED WORKS

Image processing becomes popular and rapidly grown in technologies nowadays. It is a method to convert an image such as text printed or handwritten, documents, photo and video into digital form and perform an operation in order to extract some useful information from the image.

A. *Document Identification*

Document identification is one of the areas in image processing. It is used to identify, verifying and authenticate the originality of the documents and get known the author. Research on document identification conducted using existing, improved or new techniques such as centroid detection [9], triangle model [7], watermarking [10], n-grams and histograms of words [1], statistical and Gabor features [3] and also geometric distortion [11].

There are many studies carried out to identify the document using character recognition [3] such as Arabic characters [12] and [13], Malay characters [2], isolated Farsi/Arabic characters [14], English and Arabic characters [1], Tamil script characters [15] and so on. Then, the extraction and classification process will be performed in order to get the results. Some of the studies get an accurate result but some of it has lack of resources such as the size of data from the database [2] or depending on dataset training available. Moreover, the studies only relevant to their own characters and cannot be used for others which it is limited to only one language character in a document. For example, the Arabic character recognition can only be used for documents that use the Arabic language only. If want to perform character recognition using English or Chinese characters, researchers have to develop other applications. Worse yet, when there are mixing languages within a document. However, there is existing research that been done on bilingual or multilingual languages such as [16], [17] and [3]. Thus, a good technique is needed to prevail throughout any characters of the document. Nevertheless, the document identification will be implemented more easily with finding the similarity of characters that can be found in all types of characters, namely the empty space or unoccupied space.

B. *Unoccupied space*

Unoccupied space in a document is formerly known as empty space that does not belong to others. There also previous research that refers to it as word space [18] [5], document space [18], blank space [19] and mere representative space [20]. Example of unoccupied space is shown in Figure 1 below.

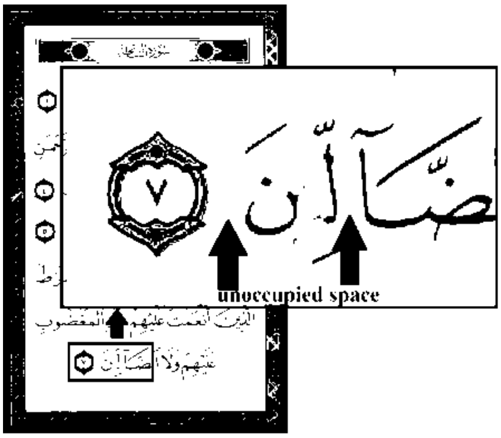

unoccupied space

Figure 1: Unoccupied spaces in a document

These unoccupied space is used for knowledge transformation [18], optical character recognition [19] and writer/author identification [20] [5]. Thus, research using unoccupied space for document identification still not exist.

C. *Thresholding*

Thresholding is an operator that select pixels that have a particular value or within a specified range. It is suitable for image interpretation. Otsu's method is one of the popular techniques of optimal thresholding. This technique used to split the image between background and object.

In [7], Otsu's method is chosen to binaries the document image which it is used to reduce the color image into the gray level image then to binary image. While [21], a binary image is a conversion of a digital image that has two values only '1' and '0'. The binary representation is a base-2 number which refers to an exclusive state in representing information. Generally, it will have two states only which are 1 and 0. When a document image is converting to binary it will distinguish the objects from its background.

Figure 2 shows the binary representation from the original image. This technique makes it possible to perform feature extraction, character recognition and identify the coordinates of each object.

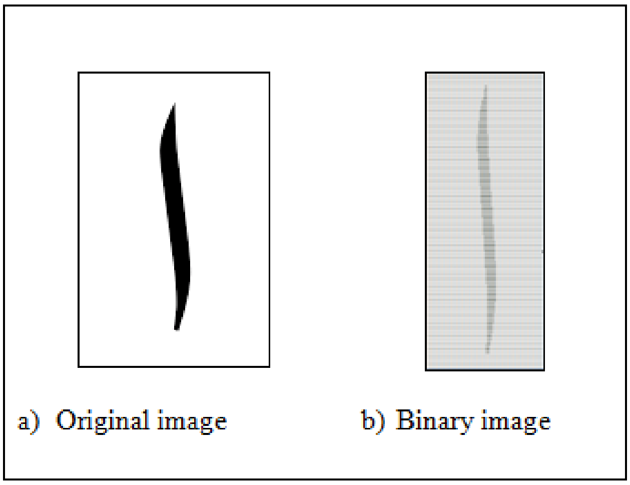


Figure 2: Conversion of original to the binary image

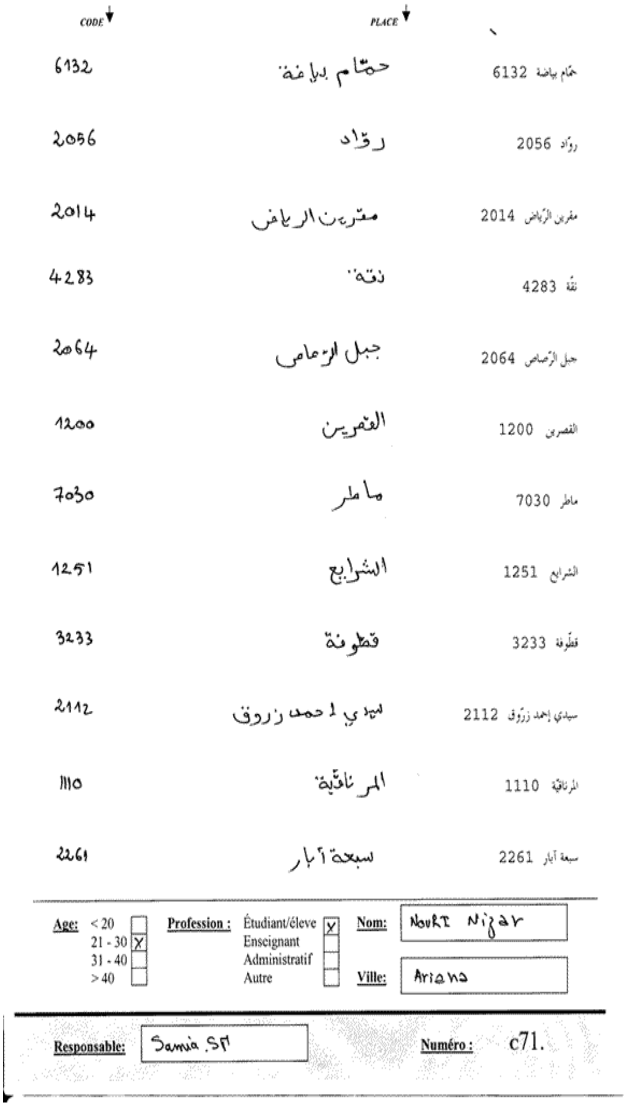
The binary image made it to possibly generate a histogram where the number of points at each level is counted by column or by row to generate a vertical and horizontal histogram.

III. Pre-Processing

The proposed method is based on finding from the previous research. There are three steps involved in the phase of data collection, binarization and histogram normalization.

A. *Data Collection*

Data collection is a method to get retrieve data and information. In this research, we used two types of data standard dataset and document image. The standard dataset which is IFN/ENIT, APTI, ACDAR AHDB, Zeki DB, AHDB/FR database is to test the effectiveness of this technique to be used with small dataset image. This dataset is handwritten Arabic from the various author. While document image is any document that been scanned such as Al-Quran. Figure 3 shows the example dataset of IFN/ENIT.



CODE	PLACE	
6132	حمام بياضة	6132 حمام بياضة
2056	رواد	2056 رواد
2014	مقرين الرياض	2014 مقرين الرياض
4283	نقة	4283 نقة
2064	جبل الرصاص	2064 جبل الرصاص
1200	القصرين	1200 القصرين
7030	ماطر	7030 ماطر
1251	الشرايع	1251 الشرايع
3233	قطوفة	3233 قطوفة
2112	سيدي احمد زروق	2112 سيدي إحمد زروق
1110	المرناقية	1110 المرناقية
2261	سبعة آبار	2261 سبعة آبار

Age: < 20 □ 21 - 30 ☒ 31 - 40 □ > 40 □

Profession : Étudiant/élève ☒ Enseignant □ Administratif □ Autre □

Nom: NouRI Nizar

Ville: Ariana

Responsable: Samia SN

Numéro : c71.

Figure 3: Sample of IFN/ENIT data set

B. Binarization

For binarization image, the data set image will be converting to binary by using the Otsu’s model method. Otsu’s model will reduce the color of the image to gray scale image. Then it will transform the image to binary in ‘1’ and ‘0’ value where the value of ‘1’ is referring to the background and ‘0’ is the object.

C. Histogram normalization

Two types of the histogram will be produced with a vertical and horizontal histogram. Vertical histogram gets the value by calculating the frequent of ‘0’ exist by a column of each pixel in the binary image. While horizontal histogram, calculate the value of ‘0’ by row.

IV. Expriments and Results

The prototype is developed using JAVA as its programming language and a folder of the document image as the database. In this research, the true experimental design was applied which there is two experiment will be executed. However, for a preliminary work only Experiment 1 will be executed. The experiment will describe as in Table 1.

Table 1
Units for Magnetic Properties

Experiment 1	
Objective	To propose a framework for feature extraction based on unoccupied space using standard dataset
Significance	To produced histogram from binary image and obtained the unoccupied space coordinates.
Input	Own dataset
Algorithm	Otsu’s model and histogram projection
Output	The coordinates list of unoccupied space is obtained from the histogram.

The experiments used own dataset image as a preliminary result. The results obtained are stored in a text file for later review. Figure 4 shows the own dataset and its binary image after binarization process using Otsu’s model.

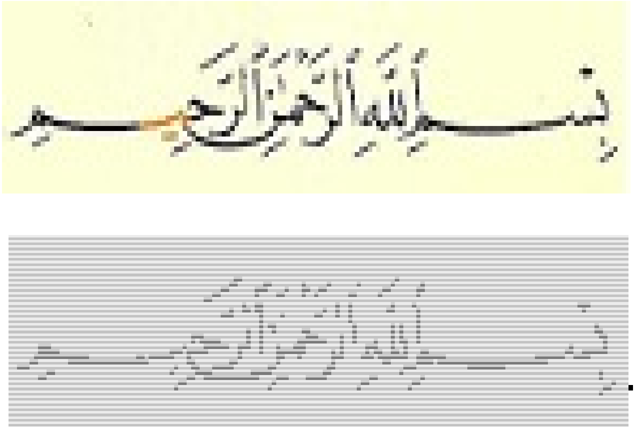


Figure 4: Own dataset and its Binary Image

Figure 5 shows the vertical histogram based on the binary image. While Figure 6 is a list of coordinates obtains from the maximum point of each column vertical histogram.

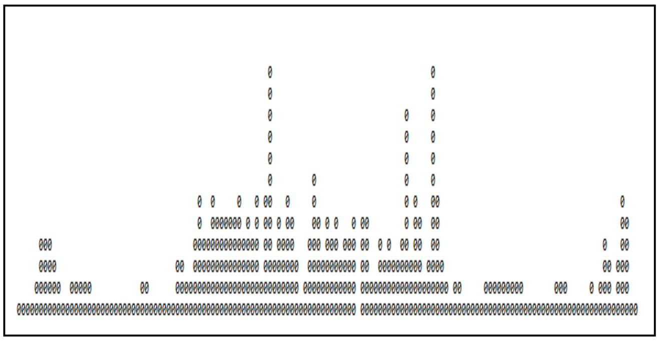
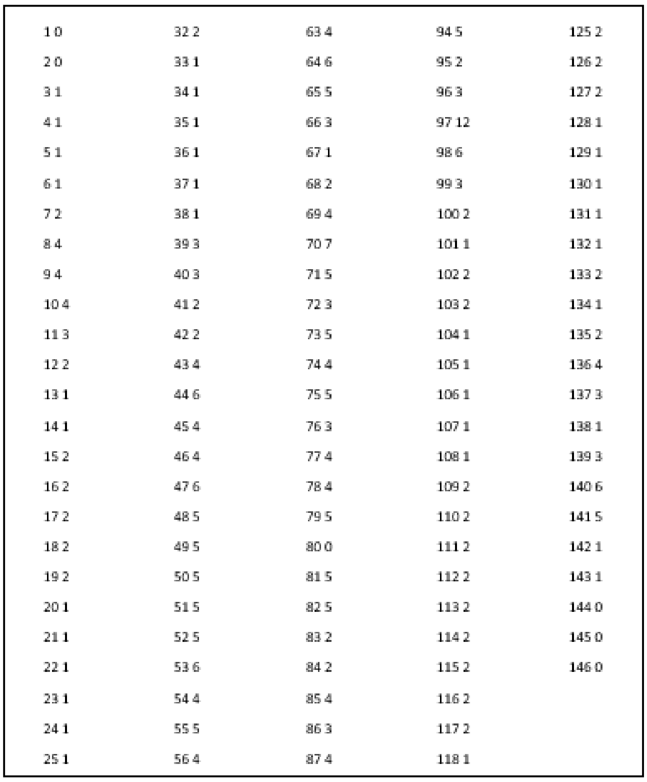


Figure 5: Vertical Histogram



1 0	32 2	63 4	94 5	125 2
2 0	33 1	64 6	95 2	126 2
3 1	34 1	65 5	96 3	127 2
4 1	35 1	66 3	97 12	128 1
5 1	36 1	67 1	98 6	129 1
6 1	37 1	68 2	99 3	130 1
7 2	38 1	69 4	100 2	131 1
8 4	39 3	70 7	101 1	132 1
9 4	40 3	71 5	102 2	133 2
10 4	41 2	72 3	103 2	134 1
11 3	42 2	73 5	104 1	135 2
12 2	43 4	74 4	105 1	136 4
13 1	44 6	75 5	106 1	137 3
14 1	45 4	76 3	107 1	138 1
15 2	46 4	77 4	108 1	139 3
16 2	47 6	78 4	109 2	140 6
17 2	48 5	79 5	110 2	141 5
18 2	49 5	80 0	111 2	142 1
19 2	50 5	81 5	112 2	143 1
20 1	51 5	82 5	113 2	144 0
21 1	52 5	83 2	114 2	145 0
22 1	53 6	84 2	115 2	146 0
23 1	54 4	85 4	116 2	
24 1	55 5	86 3	117 2	
25 1	56 4	87 4	118 1	

Figure 6: List of Coordinates from Vertical Histogram

Figure 7 shows the horizontal histogram based on the binary image and Figure 8 is a list of coordinates obtained from the maximum point of each row horizontal histogram.

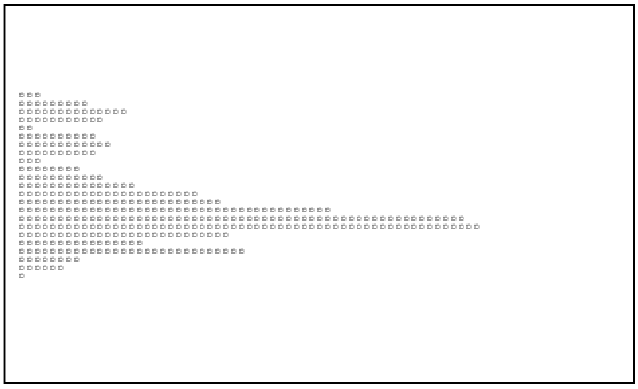
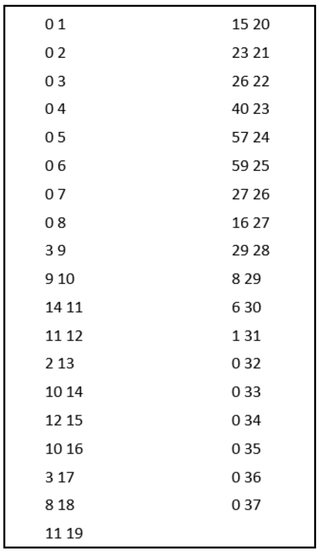


Figure 7: Horizontal Histogram



0 1	15 20
0 2	23 21
0 3	26 22
0 4	40 23
0 5	57 24
0 6	59 25
0 7	27 26
0 8	16 27
3 9	29 28
9 10	8 29
14 11	6 30
11 12	1 31
2 13	0 32
10 14	0 33
12 15	0 34
10 16	0 35
3 17	0 36
8 18	0 37
11 19	

Figure 8: List of Coordinates from Horizontal Histogram

Based on the output in Figure 6 and 8, the objective to obtain a coordinate list for both histograms is successfully retrieve and ready to be used for triangle model.

V. CONCLUSION

In this paper, a feature extraction based on unoccupied space is proposed for document identification. So far, there have not been many types of research done on unoccupied space for recognition. This research even though still in preliminary work has proven that unoccupied space can be used to recognize the identified document. The result obtained demonstrate the effectiveness of the proposed technique and readiness coordinates list for triangle model work.

ACKNOWLEDGMENT

The authors would like to express their appreciation to the Politeknik Ungku Omar for giving the study leave permission and the Ministry of Higher Education for providing the scholarship of Fellowship Training Scheme. This research also supports by grants of PJP/2016/FTMK/HI4/S01447 and FRGS/1/2017/ICT02/FTMK-CACT/F00345. Besides, thank you to the Computational Intelligence and Technologies Research Group for assisting and guiding, Faculty of Information Technology and Communication for providing the excellent research faculties and facilities and also Universiti Teknikal Malaysia Melaka for accepting author to be one of the members.

REFERENCES

[1] A. Almarimi, G. Andrejkov, and P. Sedm, “Document Verification Using n -grams and Histograms of Words,” in *IEEE 13th International Scientific Conference on Informatics · informatics’2015 · November 18-20 · Poprad · Slovakia Document*, 2015, pp. 21–26.

[2] N. Md Noh, M. R. Abdul Talib, A. Ahmad, S. A. Halim, and A. Mohamed, “Malay language document identification using,” in *Proceedings of the 10th WSEAS International Conference on NEURAL NETWORKS Malay*, 2009, no. January, pp. 163–168.

[3] S. A. Chaudhari, M. S. I. T. Programme, and R. M. Gulati, “A Comparative Analysis of Feature Extraction Techniques and Classifiers Inaccuracies for Bilingual Printed Documents (Gujarati-English),” *Int. J. Appl. Inf. Syst. – ISSN 2249-0868*, pp. 16–20, 2016.

[4] S. A. Azeem and H. Ahmed, “Effective technique for the recognition of offline Arabic handwritten words using hidden Markov models,” *Int. J. Doc. Anal. Recognit.*, vol. 16, no. 4, pp. 399–412, 2013.

[5] H. B. Barathi Ganesh, U. Reshma, and M. Anand Kumar, “Author identification based on word distribution in word space,” *2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015*, pp. 1519–1523, 2015.

[6] M. S. Azmi, “A Novel Feature From Combinations Of Triangle Geometry For Digital Jawi Paleography,” Universiti Kebangsaan Malaysia, 2013.

[7] M. S. Azmi and K. Omar, “Arabic Calligraphy Classification using Triangle Model for Digital Jawi Paleography Analysis,” in *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, 2011, pp. 704–708.

[8] N. A. Arbain, M. S. Azmi, L. B. Melhem, A. K. Muda, and H. Rashaideh, “Enhancement of Triangle Coordinate for Triangle Features for Better Classification,” *Jordanian J. Comput. Inf. Technol.*, vol. 2, no. 2, p. 107, 2016.

[9] S. H. Low, N. F. Maxemchuk, and A. M. Lapone, “Document identification for copyright protection using centroid detection,” *IEEE Trans. Commun.*, vol. 46, no. 3, pp. 372–383, 1998.

[10] R. García-soto, S. Hernández-anaya, M. Nakano-miyatake, L. Rosales-roldan, and H. Perez-meana, “Sender Verification System for Official Documents Based on Watermarking Technique,” in *10th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 2013, pp. 227–232.

[11] F. S. Joost van Beusekom, “Distortion Measurement for Automatic Document Verification,” in *2011 International Conference on Document Analysis and Recognition Distortion*, 2011, no. September, pp. 289–293.

[12] A. Lawgali, “An Evaluation of Methods for Arabic Character Recognition A.,” *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 7, no. 6, pp. 211–220, 2014.

[13] G. Abandah and M. Z. Khedher, “Analysis of Handwritten Arabic Letters Using Selected Feature Extraction Techniques,” *Int. J. Comput. Process. Lang.*, vol. 22, no. 1, pp. 1–25, 2009.

[14] A. Mowlaei, K. Faez, and A. T. Haghighat, “Feature extraction with wavelet transform for recognition of isolated handwritten farsi/Arabic characters and numerals,” in *International Conference on Digital Signal Processing, DSP*, 2002, vol. 2, pp. 923–926.

[15] K. B. Urala, A. G. Ramakrishnan, and S. Mohamed, “Recognition of open vocabulary, online handwritten pages in Tamil script,” *2014 Int. Conf. Signal Process. Commun. SPCOM 2014*, 2014.

[16] M. Mohammadi, “Parallel Document Identification using Zipf’s Law,” in *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*, 2016, no. May, pp. 21–25.

[17] A. Bozkurt, P. Duygulu, and A. E. Cetin, “Classifying fonts and calligraphy styles using complex wavelet transform,” *Signal, Image Video Process.*, vol. 9, no. 1, pp. 225–234, 2015.

[18] C. Ding, “Knowledge Transformation from Word Space to Document Space,” in *SIGIR 08*, 2008, vol. 33199, pp. 187–194.

[19] I. Baumgartner, RJ, “Blank Space Detection for Optical Character Recognition An IP . com Prior Art Database Technical Disclosure Original Publication Date : March 01 , 1973 Original Disclosure Information : TDB 03-73 p3117,” 2005.

[20] C. Djeddi, I. Siddiqi, L. Souici-Meslati, and A. Ennaji, “Codebook for writer characterization: A vocabulary of patterns or a mere representation space?,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 423–427, 2013.

[21] N. Laith, “Illumination removal and text segmnetation for Al-Quran using binary representation,” 2015.