ISSN 2278-3091
Volume 9, No.1.1, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse8191.12020.pdf
https://doi.org/10.30534/ijatcse/2020/8191.12020

Framework for document feature extraction based on unoccupied space using triangle model

A. Tahir[1], M. S. Azmi[2], N. Ahmad[2], N. A. Arbain[2] and A. R. Radzid[2]
[1]Department of Information & Communication Technology,
Politeknik Balik Pulau,
Pinang Nirai Mukim 6,
11000 Balik Pulau, Pulau Pinang, MALAYSIA
[1, 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.
azrinatahirofficial@gmail.com

ABSTRACT

Document identification is used to extract more information from documents. This information will be used to identify the document originality and to analyze whether all parts of the document written by the same author. There is research on document identification, and the problem occurred when recognition specific to one language and the size of the dataset image. Therefore, this research purposely to propose a framework for document identification which will use unoccupied space as its bases. The proposed framework adapted from the existing framework and modified to suit with this research. This research chooses document image as the document dataset. In the preprocessing phase, Otsu's method used to convert the image to binary in order to create the histogram. Then, the histogram will go through the zoning process which triangle model is used to obtain three-point coordinates for each 33 zone and produce 297 features. This features will undergo classification process where Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) as the classification techniques to obtain the result.

Key words: Document identification; unoccupied space; Otsu's method; Histogram; Triangle Model.

1. INTRODUCTION

The document is a draft or record that been written and presented in a text and graphic. Books, magazine, newspaper, Al-Quran, legal document is an example of a document which can be either online or offline. In the rapidly developing technology era, documents are readily available and accessible online; it does not only bring advantage to us but also brings disadvantages because the authenticity of a document is very dubious. Besides, the copying and modification of such a document may apply. There are also documents that the author is unknown but copied by another author.

Documents nowadays have been digitized to preserves them from aging, but plagiarism and modification exploited some of it. As a result, fake documents and the original documents mixed up through online and hard to recognize, it also becomes a big problem all over the world and causes a lack of trust from the world community. Thus, there is a need to make document identification.

The purpose of document identification is; first to ensure that the document is genuine regardless of its author and second to analyze if all parts of the document are from the same author [1].

Document identification or document recognition is one of a domain in image processing. It is used to extract useful information from documents. Most of the research using digit or character recognition in order to identify the document. However, there is a restriction on language character itself which most of the language character have their form and different among others.

The idea of this research is to identify any document regardless of the language and extract the feature using the triangle model. Since unoccupied space is a character that has similarities in all language, it will be used as the point selection to process a document and extracting its features. Then, these features go through the classification process to obtain the accurate results.

The paper has the following structure: In the second section, provides some related works of the research and its framework. The third section contains the proposed framework. The fourth section concludes the paper.

2. RELATED WORKS

Document identification is used to identify, verifying and authenticate the originality of the documents and get known the author. The purpose of document identification is to get useful information about one document as much as possible. Document identification use various recognition such as character recognition [2], signature recognition [3], digit recognition [4], handwriting recognition [5] and word space recognition [6]. Various techniques have applied such as centroid detection [7], triangle model [8], watermarking [9], n-grams and histograms of words [1] and also statistical and

Gabor features [10]. Most of them obtain an accurate result, but some of it has a lack of resources.

The size of data from the database [11], to depending on dataset training available, and the result obtained lower when test with the real document [12] is the problem faced by the researchers.

Besides, there is research that relevant to their purpose and cannot use by others because it only applies to its language, digit or sub-word in a document such as Arabic [13] [14], Malay [11], Tamil [15] and English [1]. For example, the character recognition for Arabic only applicable for documents with the Arabic language.

The problem becomes worst when a document consists of more than one language because each language has a different and unique character on its own. However, there is research done on bilingual or multilingual languages such as [16] and [17].

There are several frameworks proposed by other research in document identification that is equivalent to this research like Arabic character recognition framework [18], keyword retrieval system framework [19] and Digital Jawi Palaeography framework [20].

2.1. Arabic character recognition framework

The Arabic character recognition framework developed by Jafaar Al Abodi and Xue Li (2014). This framework has three primary processes: pre-processing, segmentation and features extraction.

Pre-processing phase has two stages binarization and skeletonization. Binarization uses a new method that applies double-threshold based on the image edge and the text intensity. While skeletonization, reduce the width of a pattern shape to a single pixel.

Segmentation phase, a new technique that based on the geometrical features of Arabic handwriting characters and independent from it style and size. In the features extraction phase, to preserve the cursive writing properties a new coding scheme to record the geometrical features of pixels.

2.2. Keyword retrieval system framework

Hongxi and Guanglai [19], proposes a keyword retrieval system framework consist of the offline and online document. In the offline part, the scanned document images converted into binary images. Then, word images obtained by connected components analysis on each binary image which will be used to extract several profile features for each word image.

Discrete Fourier transform (DFT) performed on each profile feature, and the appropriate number of the DFT lower-order coefficients is reserved to form a fixed-length feature vector for online image-to-image matching. Finally, both the DFT coefficients and the coordinate information of word images saved into a database. While the online part, the user provides a query word image by a synthesis tool.

The processes of the feature extraction and the fixed-length representation performed as same as in the off-line part. Moreover, the ranked results returned by sorting similarities between the query word image and each word image in the database.

2.3. Digital Jawi Palaeography framework

Digital Jawi Paleography framework is a framework proposed by Mohd Sanusi Azmi (2011). This framework divided into three phase: Data Collection and Preprocessing, Triangle Model and Zoning Features and Classification and Evaluation.

In the first phase, the data collection process involves the use of two types of datasets. Dataset types are both standard and local datasets. Local data set, involves the development of data sets written by khat writers, then scanned and segmented.

The images of the khat which have segmented and made by naming before entering the pre-processing process. Standard and local dataset images in the pre-processing process will be converted to a binary image using Otsu's Method and performed labeling.

In the second phase, a new technique which is the triangle model. This technique uses three-point selection for each 33 zone produced to extract vector features.

The classification is carried out to obtain percent recognition accuracy. The research using a different type of testing Unsupervised Machine Learning and Supervised Machine Learning. Seven algorithms for Unsupervised Machine Learning have been used to classify Arabic calligraphy which is Euclidean, Manhattan, Chebyshev, Minkowski, Sorenson, Correlation and Angular Separation. While Supervised Machine Learning using SVM classification to obtain percentage accuracy.

3. PROPOSED FRAMEWORK

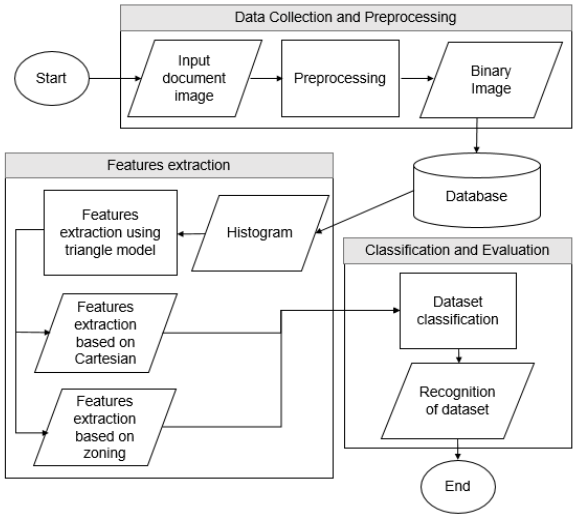


Figure 1: Proposed Framework

The proposed framework derived from previous research with some modifications to suit the research work. The framework divided into three phase: Data Collection and preprocessing, Features extraction and the third phase is Classification and Evaluation.

3.1. Data collection and pre-processing

The document image used for data collection and can be collected from various standard datasets such as Tobacco-800, Impact and PRIma dataset. This research chooses To-

bacco-800 as the document image dataset. After the dataset is determined, the pre-processing process will take part.

Otsu's method is the technique selected to perform the pre-processing process. Otsu's method is one of the favorite techniques in optimal thresholding. This technique uses to distinguish between backgrounds and object by reducing the color image into gray level image and convert it to a binary image. The binary image consists of binary representation '1' and '0' which is '1' refer to the background and '0' refer to object.

The resulting binary image needs to eliminate the unoccupied space that exists around the words. This unoccupied space refers to unoccupied space of the document, not unoccupied space that exists between words as in Figure 2.

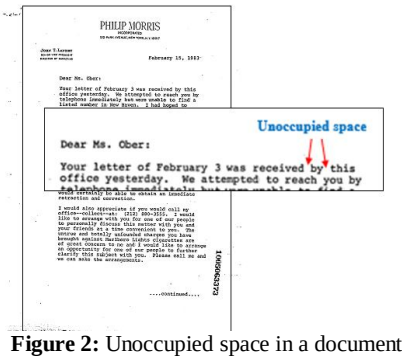
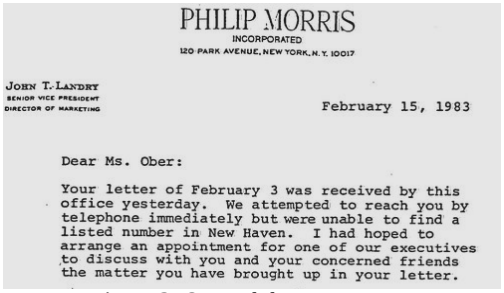


Figure 2: Unoccupied space in a document

Alia and Mustafa (2016) [21], eliminate unoccupied space around the words in an image because it does not have meaning in any recognition process. The eliminating process use bounding box which forms from four points of the first pixel of the written words. The image used in this process is a binary image whereas background and object are distinguished. Figure 3 to 6 show the image process.


PHILIP MORRIS
INCORPORATED
120 PARK AVENUE, NEW YORK, N.Y. 10017

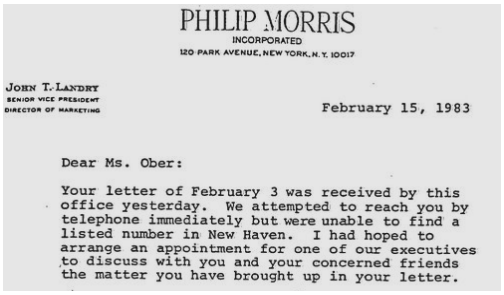
JOHN T. LANDRY
SENIOR VICE PRESIDENT
DIRECTOR OF MARKETING

February 15, 1983

Dear Ms. Ober:

Your letter of February 3 was received by this office yesterday. We attempted to reach you by telephone immediately but were unable to find a listed number in New Haven. I had hoped to arrange an appointment for one of our executives to discuss with you and your concerned friends the matter you have brought up in your letter.

Figure 3: Original document image


PHILIP MORRIS
INCORPORATED
120 PARK AVENUE, NEW YORK, N.Y. 10017

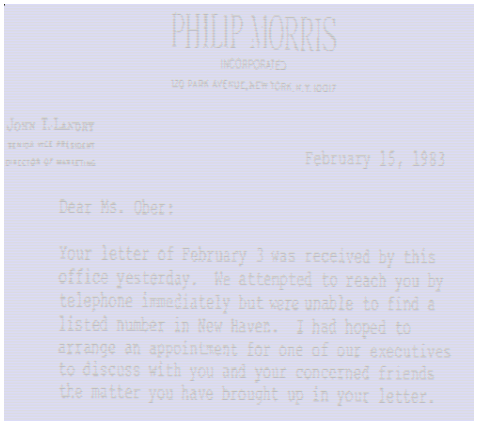
JOHN T. LANDRY
SENIOR VICE PRESIDENT
DIRECTOR OF MARKETING

February 15, 1983

Dear Ms. Ober:

Your letter of February 3 was received by this office yesterday. We attempted to reach you by telephone immediately but were unable to find a listed number in New Haven. I had hoped to arrange an appointment for one of our executives to discuss with you and your concerned friends the matter you have brought up in your letter.

Figure 4: Gray level document image


PHILIP MORRIS
INCORPORATED
120 PARK AVENUE, NEW YORK, N.Y. 10017

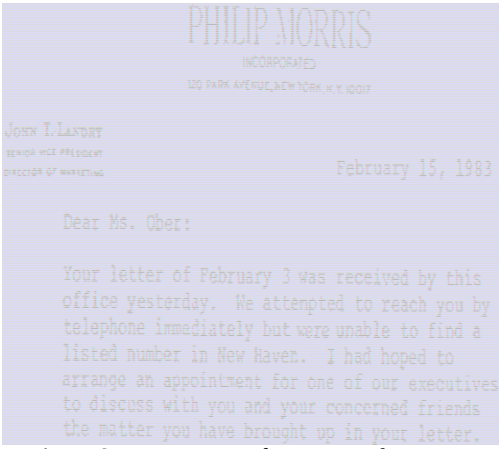
JOHN T. LANDRY
SENIOR VICE PRESIDENT
DIRECTOR OF MARKETING

February 15, 1983

Dear Ms. Ober:

Your letter of February 3 was received by this office yesterday. We attempted to reach you by telephone immediately but were unable to find a listed number in New Haven. I had hoped to arrange an appointment for one of our executives to discuss with you and your concerned friends the matter you have brought up in your letter.

Figure 5: Binary image


PHILIP MORRIS
INCORPORATED
120 PARK AVENUE, NEW YORK, N.Y. 10017

JOHN T. LANDRY
SENIOR VICE PRESIDENT
DIRECTOR OF MARKETING

February 15, 1983

Dear Ms. Ober:

Your letter of February 3 was received by this office yesterday. We attempted to reach you by telephone immediately but were unable to find a listed number in New Haven. I had hoped to arrange an appointment for one of our executives to discuss with you and your concerned friends the matter you have brought up in your letter.

Figure 6: Binary image after process elimination

3.2. Features extraction

At this phase, the resulting binary image will be analyzed and form to the vertical histogram. The analysis on binary image done by calculates the occurrence of value '0' that represent to object. After the calculation process, it will create a histogram based on the occurrence pixel value as shown in Figure 7.

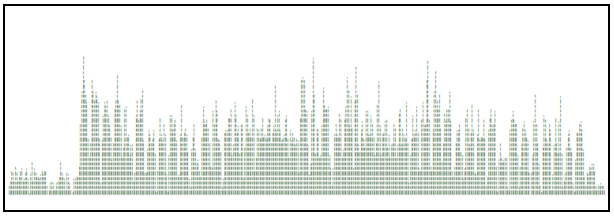


Figure 7: Vertical histogram

Then, the resulting histogram will process features extraction using the combination of triangle model and zoning. The triangle model uses two features vector which is Cartesian and Zoning to obtained three-point coordinates that will form a triangle for each 33 zone and produce 297 features.

Table 1: Zoning and features produce.

No.	Zone type	Quantity zone	Quantity features
1	Cartesian zoning	5	45
2	Horizontal zoning	6	54
3	Direkayasa vertical zoning	14	126
4	45 degree zoning	8	72
		Total	297

3.3. Classification and evaluation

Classification and evaluation phase will use the features extraction result with Random Forest and SVM classifier to get the percent recognition accuracy and compare among each other. The highest classification result use as the standard for proposed features.

4. CONCLUSION

As the conclusion, this paper proposed a framework for document features extraction based on unoccupied space using triangle model. The research focuses on features extraction by using the triangle model. Compare to previous research that chooses digit or character recognition, this research use unoccupied space as the pattern recognition.

Otsu's method is used to produce the binary image which process elimination will take part to eliminate the unwanted unoccupied space of the document. A histogram was produced from the calculation of occurrence object pixel value which is '0'. This histogram distinguishes words and unoccupied space that exist between words. In this process, a histogram as the input image will be used to extract features. The combination of the triangle model and features vector: Cartesian and zoning produce the features extraction. These features use the Random Forest and SVM classifier to get the recognition accuracy result and compare each result. The highest classification is the final result recognition. The final result will be compared to others research in order to test the effectiveness of this research.

ACKNOWLEDGMENT

Authors want to give an appreciation to the Politeknik Balik Pulau and the Ministry of Higher Education for providing the scholarship of Fellowship Training Scheme. This research also supports by grants of FRGS/1/2017/ICT02/FTMK-CACT/F00345. Faculty of Information Technology and Communication and Universiti Teknikal Malaysia Melaka for providing the excellent research faculties and facilities.

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] 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.

[3] S. Ahmed, M. I. Malik, M. Liwicki, and A. Dengel, "Signature segmentation from document images," *Proc. - Int. Work. Front. Handwrit. Recognition, IWFHR*, pp. 425–429, 2012. https://doi.org/10.1109/ICFHR.2012.271

[4] 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.

[5] Z. Razak *et al.*, "Off-line Jawi handwriting recognition using hamming classification," *Inf. Technol. J.*, vol. 8, no. 7, pp. 971–981, 2009.

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

[7] 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. https://doi.org/10.1109/26.662643

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] T. Marušić, Ž. Marušić, and Ž. Šeremet, "Identification of authors of documents based on offline signature recognition," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, no. May, pp. 1144–1149, 2015.

[13] 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.

[14] G. Abandah and M. Z. Khedher, "Analysis of Handwritten Arabic Letters Using Selected Feature Extraction Techniques," *Int. J. Comput. Process. Lang.*, vol. 22, no. 01, pp. 1–25, 2009. https://doi.org/10.1142/S1793840609001981

[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] J. Al Abodi and X. Li, "An effective approach to offline Arabic handwriting recognition," *Comput. Electr. Eng.*, vol. 40, no. 6, pp. 1883–1901, 2014.

[19] H. Wei and G. Gao, "A keyword retrieval system for historical Mongolian document images," *Int. J. Doc. Anal. Recognit.*, vol. 17, no. 1, pp. 33–45, 2014.
[20] M. S. Azmi, "A Novel Feature From Combinations Of Triangle Geometry For Digital Jawi Paleography," Universiti Kebangsaan Malaysia, 2013.
[21] M. S. Kadhm and A. Karim Abdul Hassan, "Arabic handwriting text recognition based on efficient segmentation, DCT and HOG features," *Int. J. Multimed. Ubiquitous Eng.*, vol. 11, no. 10, pp. 83–92, 2016. https://doi.org/10.14257/ijmue.2016.11.10.07