

A New Adaptive Method for Extracting Header Words from Official Printed Arabic Documents

Assist. Prof. Dr. Matheel E. Abdulmunim
Computer Science Department
University of Technology
Baghdad, Iraq
matheel_74@yahoo.com

Haithem K. Abass
Software Engineering and Information
Technology Department
Al-Mansour University College
Baghdad, Iraq
haithem_72@yahoo.com

Abstract

Words extraction techniques from documents have very significant and effective role in document image analysis and retrieval systems. In this paper, a new method has been proposed for detecting and extracting header words from official printed Arabic documents. In the proposed method line of Arabic words with various fonts, styles, and sizes have been extracted from printed Arabic documents with different shapes, colors and resolutions. The extraction of header words based on effective segmentation technique that will separate different objects in a document including text lines, graphics, signature, logo, and other objects. The segmentation operation depends on document analysis that will efficiently predict vertical and horizontal distances between objects in Arabic documents. After segmentation operation, header words detection will performed by using sequence of influential rules within decision tree that correctly detected header words in a document image. Finally, list of header words will extracted as separated text lines from document image. Extracted header words can be utilized in many applications like words matching, words spotting, documents

classification, documents retrieval and other applications that depends on words extraction. In this paper, a dataset of different official printed Arabic documents has been constructed and tested by the proposed method. These Arabic documents dataset obtained and gathered from various official institutions websites and offices. The proposed Arabic header words extraction method obtained 96% for recall, 98% for precision and 97% for f-score.

Keywords: document segmentation, run length smearing, header words, words extraction

طريقة جديدة ومتكيفة لاستخراج كلمات الرأس من الوثائق العربية الرسمية المطبوعة

الخلاصة

أصبحت لتقنيات استخراج الكلمات من الوثائق دور مهم ومؤثر في أنظمة تحليل واسترجاع الوثائق المصورة. تم في هذا البحث اقتراح طريقة جديدة لتحديد واستخراج كلمات الرأس من الوثائق العربية الرسمية المطبوعة. تم في هذه الطريقة استخراج عبارات من الكلمات العربية متنوعة الخطوط والأنماط والأحجام من الوثائق العربية المطبوعة المختلفة الأشكال والألوان والدقة. عملية استخراج كلمات الرأس تعتمد على تقنية تجزئة فعالة تعمل على فصل مكونات الوثائق المتضمنة النصوص والشعارات والرسومات والتواقيع وغيرها. عملية التجزئة تعتمد على تحليل الوثيقة والتي

يمكن من خلالها استنتاج ابعاد المسافات الافقية والعمودية بين المكونات. بعد عملية التجزئة يتم تحديد كلمات الرأس من خلال سلسلة من القواعد المؤثرة مع شجرة اتخاذ القرار التي سوف تحدد بشكل صحيح كلمات الرأس في الوثيقة المصورة. الكلمات المستخلصة يمكن الانتفاع منها في الكثير من التطبيقات مثل مطابقة الكلمات، اكتشاف الكلمات، تصنيف واسترجع الوثائق وغيرها من التطبيقات التي تعتمد على استخراج الكلمات. تم في هذا البحث بناء مجموعة بيانات من وثائق عربية رسمية مطبوعة واختبارها في الطريقة المقترحة. هذه الوثائق العربية تم الحصول عليها وتجميعها من مختلف المواقع الالكترونية الرسمية ومن المكاتب. الطريقة المقترحة لاستخراج كلمات الرأس من الوثائق العربية حصلت على 96% لنسبة الاستدعاء و98% لنسبة الدقة و97% لمعامل الهدف.

كلمات المفتاح: تجزئة الوثيقة، شجرة القرار، كلمات الرأس، استخراج الكلمات.

Arabic fonts with different sizes and styles by using Run Length Smearing Algorithm (RLSA) technique. Table 1 shows samples of various Arabic fonts used in the constructed documents dataset. Segmentation of Arabic words is difficult than other languages like European languages because of its nature that contains separated, connected and overlapped characters. These difficulties in segmentation can overcome by analyzing documents and estimate the average vertical and horizontal distances between objects especially words and lines.

Table 1. Samples of Various Arabic Fonts.

Arabic Words	Type of Font
مجموعة بيانات لوثائق عربية رسمية مطبوعة مصورة	Traditional Arabic
مجموعة بيانات لوثائق عربية رسمية مطبوعة مصورة	Arial
مجموعة بيانات لوثائق عربية رسمية مطبوعة مصورة	Calibri
مجموعة بيانات لوثائق عربية رسمية مطبوعة مصورة	Simplified Arabic
مجموعة بيانات لوثائق عربية رسمية مطبوعة مصورة	Mudir MT
مجموعة بيانات لوثائق عربية رسمية مطبوعة مصورة	Times New Roman

The rest of this paper is organized as follows. In section 2, the related work has been reviewed. In section 3, the proposed header words extraction technique is illustrated. In section 4, experimental results are evaluated and discussed and finally section 5, presents the conclusions of the proposed technique.

1) INTRODUCTION

Many document images, which include text, contain significant and meaningful information that can be used to obtain a good notification about these documents. Extracting important part from text images is very useful in many applications like document analysis and indexing, information retrieval, digital library, and text summarization [1]. A document image may contains text, tables, graphical shapes and other objects with different colors and resolutions. Therefore, extracting important and useful words is critical and challenge problem [2]. This paper propose a segmentation and extraction technique for printed Arabic documents images. The presented technique attempt to extract sequence of words from header documents located in multiple text lines. Segmented header words may define important information that can be used to know the origination of documents. This property is very useful and crucial in documents classification, recognition and retrieval operations. Effective segmentation technique is applied to Arabic documents dataset that contain variety of

2) RELATED WORK

In the proposed technique, text line segmentation is implemented to extract sequence of words as a line from the header of

scanned documents. Many researches that developed and enhanced text line segmentation techniques for Arabic or semi Arabic documents in recent years. A. Zahour et al. [3] propose handwritten line segmentation by using Horizontal Projection Profile (HPP) technique. Jayant Kumar et al. [4] use graph-based approach for handwritten text lines segmentation. Z. Shi. [5] developed technique for line segmentation depended on a generalized Adaptive Local Connectivity Map (ALCM). M. Khayyat et al. [6] propose dilation method with a dynamic adaptive mask to extract lines from Arabic document image. M. A. Mohammed [7] introduce a segmentation method for text line handwritten for based on line height scheme. A. Al-dmour and F. Fraij [8] present new technique for segmentation Arabic documents into words and lines base on autocorrelation to enhance the self-similarity of features. M. Younes, and Y. Abdellahb [9] developed three different approaches of Arabic line segmentation, and compared between these approaches. N. Aouadi and A. K. Echi [10] propose a method for text line segmentation and recognition by using Markovian classifier. Y. Bouldid et al. [11] present an approach for Arabic manuscripts lines segmentation based on multi agent method. B. Biswas et al. [12] modify a robust scheme for lines segmentation based on divide and conquer strategy.

3) PROPOSED HEADER WORDS EXTRACTION MODEL

A block diagram of the proposed header words extracting model illustrate in figure (1). This structure consists of a number of steps each of which performs a specific task in the extraction process.

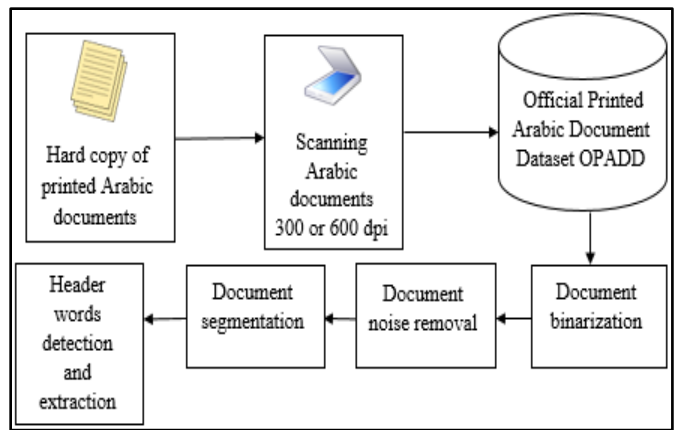


Figure 1. Block diagram for Arabic header words extraction.

A. DATASET CONSTRUCTION

An Official Printed Arabic Document Dataset (OPADD) has been constructed and tested in the proposed segmentation model. This dataset consist of different forms of official printed Arabic document images obtained from various authorized web sites like ministries, universities, government institutions and other official states. The dataset represents various types of Arabic documents like letters, reports, books, forms, announcements, administrative instructions and other official documents. All these papers should have Arabic header words and may contain other objects such as texts, logos, borders, graphics, and other objects. These documents may stored in printable format as Portable Document Format (PDF), or Microsoft word document (DOC). The pages of Arabic documents are printed using HP Deskjet printer 2540 series and then scanned by scanner with 300dpi and 600dpi resolutions. More than 300 pages are printed and scanned in portrait orientation and in landscape orientation. These scanned document images are stored in three types of color level, first level is black and white image of 1-bit per pixel, second level is grayscale image of 8-bits per pixel, and third level is true color image of 24-





bit per pixel. After scan operation, each document image in dataset is stored as a file of JPG (Joint Photographic Group) file format. Table 2 shows the classification of constructed Arabic documents dataset according to their color and resolutions.

Table 2. Distribution of document images in the dataset.

Class	Image Color and Resolution						Total
	True Color		GrayScale		Black and White		
	300 dpi	600 dpi	300 dpi	600 dpi	300 dpi	600 dpi	
Report	12	8	10	4	10	6	50
Book	18	6	16	6	8	2	56
Letter	20	8	14	4	15	3	64
Form	8	2	10	4	8	2	34
Announcement	12	4	8	2	14	6	46
Instruction	16	6	14	4	8	4	52
Total	86	34	72	24	63	23	302

Samples of document images of different categories shown in Table 3.

Table 3. Samples of document images in the dataset.

Class	Image Sample	Class	Image Sample
Report		Form	
Book		Announcement	



B. PREPROCESSING

Preprocessing involves binarization as a first preprocessing step that will convert any color or gray scale document image $f(x,y)$ into black and white image $bw(x,y)$ according to the Local Threshold value LT computed depending on the value of pixels in a document image as shown below:

$$\text{If } f(x,y) \geq LT \text{ then } bw(x,y)=1$$

$$\text{Otherwise } bw(x,y)=0$$

Noise removal is the second preprocessing step that will enhance the document images. Using accurate and appropriate value of LT from previous step that will greatly reduce some type of noises like small gaps and can sharp edges of the objects. Salt and paper noise will be removed by using median filter. In addition, unnecessary objects like dark margins and some border lines can also be detected and eliminated without effect or degrade logos.

C. DOCUMENT SEGMENTATION

For line level segmentation, run length smearing method has been modified and applied to segment document images. The modifications comes from that RLSA technique is applied in horizontal and vertical directions with different variable threshold values and with constant factor that will enhance the smearing operation. These threshold values are computed to control the number of sequence of pixels that will be smeared in the image. For the proposed approach, bounding box is constructed for each connected component in a

binary document and then histogram is computed and estimate the value of smeared threshold value L . To achieve best results, the value of L is multiply by a constant factor Ch for horizontal smearing, and Cv for vertical smearing. The final smeared and segmented image $sm(x,y)$ is produce by performing logical AND operation between horizontal smeared image $hsmage(x,y)$ and vertical smeared image $vsimage(x,y)$. Algorithm 1 presents the proposed segmentation technique.

Algorithm 1: modified run length smearing

Input: Binary document image $g(x,y)$.

Output: smeared segmented image $s(x,y)$.

Begin

Step 1: Compute connected component of $g(x,y)$.

Step 2: Find bounding box for connected component resulting $b(x,y)$.

Step 3: Calculate image histogram for bounding box to estimate the value of smearing threshold value L .

Step 4: Apply horizontal RLS to $b(x,y)$ with $limit_smear_value = L * Ch$ resulting image $hsmage(x,y)$.

Step 5: Apply vertical RLS to $b(x,y)$ with $limit_smear_value = L * Cv$ resulting image $vsimage(x,y)$.

Step 6: Perform AND operation between image $hsmage(x,y)$ and $vsimage(x,y)$ resulting final smeared image $s(x,y)$.

End

Figure (2) shows the main steps of segmentation algorithm with horizontal and vertical smearing operations for a portion of Arabic document. In horizontal smearing, a number of words with their components are

merged together as a black region. In vertical smearing a number of Arabic characters points are combined with their related characters in the words. Making logical AND between horizontal and vertical smeared objects will eliminate gaps between them.

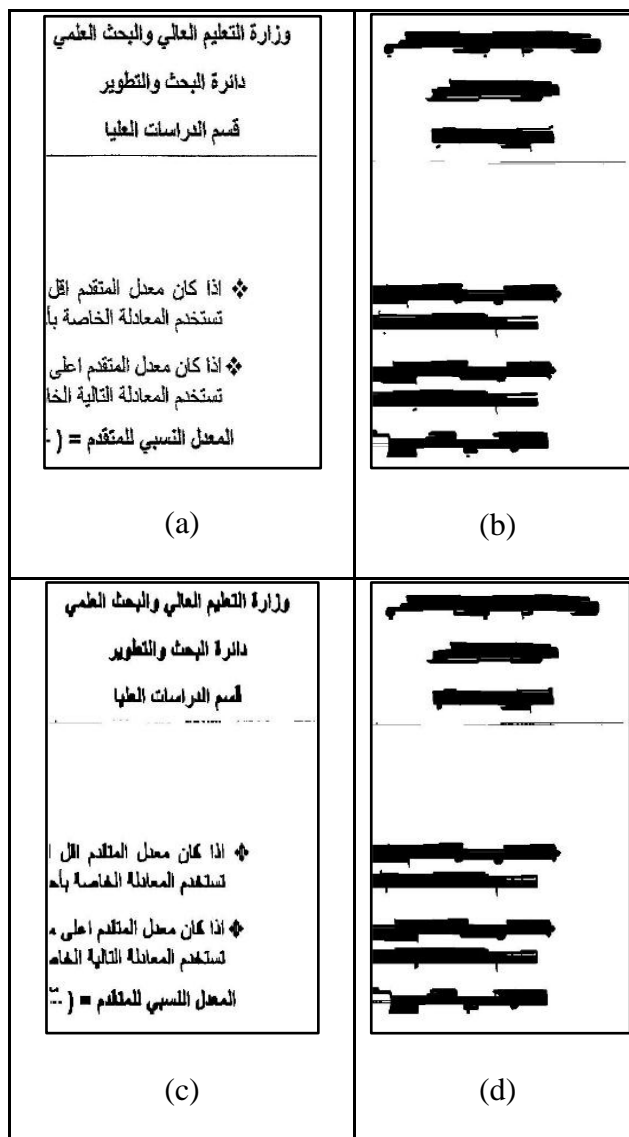


Figure 2. Document segmentation using smearing RLSA technique. (a) Binary image; (b) Image after horizontal smearing; (c) Image after vertical smearing; (d) Image after logical AND between horizontal and vertical smeared images.

D. HEADER WORDS EXTRACTION

After segmentation by using smearing method, each region in a document image is

represented as a bounding block with specified **X** and **Y** coordinates and dimensions **W** for width and **H** for height. Significant and appropriate features related to the bounding rectangles are extracted and saved. These features are applied as rules in the decision tree to correctly detect the header words and extract them as separated text lines from the document. In the proposed method, block features are computed for each bounding box according to the following equations:

1) **Y_pos**: y coordinate of block (**Y**) relative to the height of image (**height**).

$$Y_pos = Y / \text{height} \dots\dots\dots (1)$$

2) **Width_ratio**: width of block (**W**) relative to the width of image (**width**).

$$\text{Width_ratio} = W / \text{width} \dots\dots\dots (2)$$

3) **Height_ratio**: height of block (**H**) relative to the height of image (**height**).

$$\text{Height_ratio} = H / \text{height} \dots\dots\dots (3)$$

4) **Block_density**: number of foreground pixels in the block (**N**) relative to the area of block (**area**).

$$\text{Block_density} = N / \text{area} \dots\dots\dots (4)$$

Constructing a decision tree consisting of four nodes as shown in figure (3) will support accurate decision for extracting and distinguishing segmented header words from other objects like logos and borders. These nodes represent suitable and sufficient rules that will applied to test block features obtained previously with appropriate range. The sequence of these nodes in a tree is implemented according to their significance and weight in making header words decision.

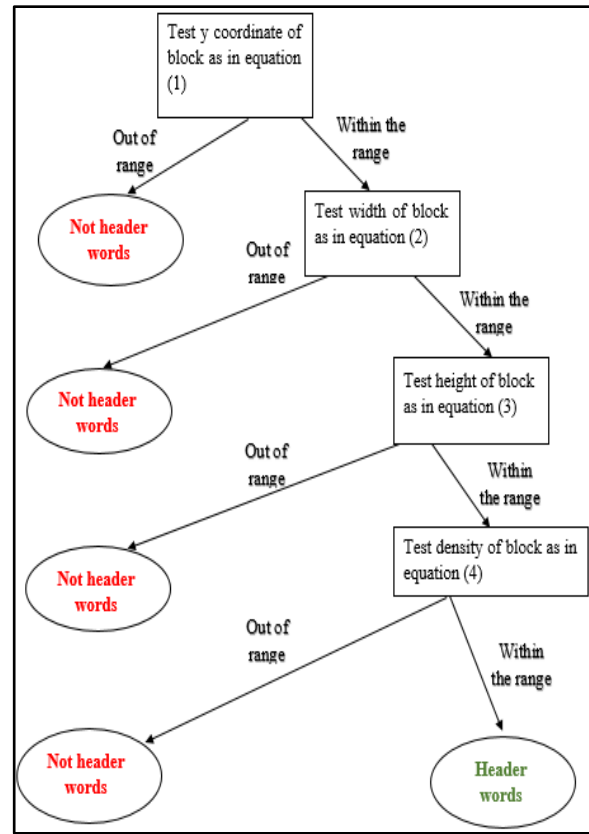


Figure 3. Decision tree to detect header words from different segmented blocks.

Feature equations for decision tree nodes and their suitable ranges are illustrated in table 4.

Table 4. Appropriate range values of header words for each node.






Node Number	Feature Equation	Appropriate Range for Header Words
Node 1	Equation (1)	[0 , 0.20]
Node 2	Equation (2)	[0.060 , 0.45]
Node 3	Equation (3)	[0.0035 , 0.030]
Node 4	Equation (4)	[0.25 , 0.65]



4) EXPERIMENTAL RESULTS

The proposed extraction technique has been tested with more than 300 of real and official Arabic documents of different types,

colors, and resolutions. The tested documents contain Arabic text of various fonts and other objects like logos, stamps, signatures, tables, graphics, and English words. The extraction technique is able to detect and extract a number of header words from each Arabic document image in the dataset. Table 5 shows different input image samples and their related header words which are extracted using the proposed technique.

Table 5. Samples of document images and their related header words.

Input Image	Extracted Header Word
	جمهورية العراق
	وزارة الداخلية
	مديرية الجنسية العامة
	مديرية شؤون البطاقة الوطنية
	جمهورية العراق
	وزارة التربية
	المديرية العامة للمناهج
	جمهورية العراق
	وزارة التعليم العالي والبحث العلمي
	جهاز الاشراف والتقويم العلمي
	وزارة التعليم العالي والبحث العلمي
	المجلس العراقي للاختصاصات الطبية
	جمهورية العراق
	وزارة الداخلية
	مديرية الاحوال المدنية والجوازات والأقامة

	جمهورية العراق
	وزارة الصحة
	دائرة الأمور الفنية
	مكتب الاعلام الدوائي
	جمهورية العراق
	هيئة النزاهة
	دائرة الوقاية

As shown in the above table, most of Arabic header words are detected in the top right side of document image and they consist of 2 to 4 text line. Each text line contains 2 to 5 words. The performance of the proposed extraction technique is evaluated by taken into account the following criteria recall, precision, and f-score as shown in the following equations:

$$\text{Recall} = \frac{\text{number of header words correctly extracted}}{\text{number of actual header words}} \dots\dots\dots (5)$$

$$\text{Precision} = \frac{\text{number of header words correctly extracted}}{\text{number of extracted words}} \dots\dots (6)$$

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \dots\dots\dots (7)$$

According to this evaluation of performance, the proposed Arabic header words extraction technique obtained 96% for recall, 98% for precision and 97% for f-score.

5) CONCLUSIONS

The purpose of this paper is to develop adaptive Arabic header words extraction method that detects and extracts a number of words from the header part of document image. Header words contains significant and important information that define the origin of the document. The extraction method has been tested by using constructed dataset named

OPADD consisting of more than 300 printed official Arabic documents. The dataset is real and contains different categories of documents like reports, letters, forms, announcements, book pages and other official papers. For document segmentation, modified horizontal plus vertical smearing algorithm has been used for line level segmentation. For header words detection and extraction, effective decision tree has been implemented to correctly distinguish header words from other objects. Extraction Arabic header words is very useful in many document image processing applications and can be considered as an important step for document image analysis, recognition, classification and retrieval.

REFERENCES

- [1] L. A. Wang, W. Fan, J. Sun, S. Naoi, T. Hiroshi, "Text Line Extraction in Document Images", 13th International Conference on Document Analysis and Recognition, pp. 191-195, 2015.
- [2] H. Ghaleb, P. Nagabhushan, and U. Pal, "Graph Modeling based Segmentation of Handwritten Arabic Text into Constituent Subwords," *Int. J. Image, Graph. Signal Process.*, vol. 8, no. 12, pp. 8–20, 2016.
- [3] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic hand-written text-line extraction," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 281–285, 2001.
- [4] Kumar, J., Abd-Almageed, W., Kang, L., And Doermann, D, "Handwritten arabic text line segmentation using affinity propagation," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS)*. pp. 135–142, 2010.
- [5] Z. Shi, S. Setlur, and V. Govindaraju, "A steerable directional local profile technique for extraction of handwritten Arabic text lines," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, no. i, pp. 176–180, 2009.
- [6] M. Khayyat, L. Lam, C. Y. Suen, F. Yin, and C. L. Liu, "Arabic handwritten text line extraction by applying an adaptive mask to morphological dilation," in *Proceedings - 10th IAPR International Workshop on Document Analysis Systems, DAS 2012*, pp. 100–104, 2012.
- [7] M. A. Mohammed, M. R. Kumar, and R. Pradeep, "Text Line Segmentation of Arabic Handwritten Documents using Line Height Method," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. 11, pp. 170–174, 2014.
- [8] A. Al-dmour and F. Fraij, "Segmenting Arabic Handwritten Documents into Text lines and Words," *International Journal of Advancements in Computing Technology(IJACT)*, vol. 6, no. 3, pp. 109–119, 2014.
- [9] M. Younes and Y. Abdellah, "Segmentation of Arabic Handwritten Text to Lines," in *Procedia Computer Science*, vol. 73, no. Awict, pp. 115–121, 2015.
- [10] N. Aouadi and A. K. Echi, "Word Extraction and Recognition in Arabic Handwritten Text," vol. 12, no. 1, pp. 17–23, 2016.
- [11] Y. Boulid, A. Souhar, and M. Y. Elkettani, "Segmentation approach of Arabic manuscripts text lines based on multi agent systems," *International Journal of Computer Information Systems and Industrial Management Applications*. Vol. 8, pp. 173-183 March, 2016.
- [12] B. Biswas, U. Bhattacharya, and B. B.

Chaudhuri, "A Robust Scheme for Extraction of Text Lines from Handwritten Documents," in Proceedings of International Conference on Computer Vision and Image Processing. vol. 2, pp. 107–116, 2017.