DOI: 10.24996/ ijs.2017.58.3C.18

Iraqi Journal of Science

ISSN: 0067-2904

Header-Words Based for Printed Arabic Document Images Retrieval System

Matheel E. Abdulmunim[1], Haithem K. Abass[*2]

[1] Department of Computer Science, University of Technology, Baghdad, Iraq.

[2]Software Engineering and InformationTechnology Department, Al Mansour University College, Baghdad, Iraq.

Abstract

Printed Arabic document image retrieval is a very important and needed system for many companies, governments and various users. In this paper, a printed Arabic document images retrieval system based on spotting the header words of official Arabic documents is proposed. The proposed system uses an efficient segmentation, preprocessing methods and an accurate proposed feature extraction method in order to prepare the document for classification process. Besides that, Support Vector Machine (SVM) is used for classification. The experiments show the system achieved best results of accuracy that is 96.8% by using polynomial kernel of SVM classifier.

Keywords: DIR, Segmentation, Header-words, Words spotting, SVM.

نظام لاسترجاع الوثائق العربية المطبوعة بالاعتماد على كلمات الرأس

مثيل عماد الدين عبد المنعم[1]، هيثم كريم عباس[2*]

قسم علوم الحاسوب، الجامعة التكنولوجية، بغداد، العراق.

قسم هندسة البرمجيات وتكنولوجيا المعلومات، كلية المنصور الجامعة، بغداد، العراق.

الخلاصة

أنظمة استرجاع الوثائق العربية المطوعة لها دور مهم وضروري في الشركات والحكومات ومختلف الاستخدامات. تم في هذا البحث اقتراح نظام الاسترجاع الوثائق العربية الرسمية المصورة بالاعتماد على اكتشاف كلمات الراس. النظام المقترح يستخدم طريقة كفوؤة في تجزئة الوثاق والمعالجة الأولية لها وطريقة دقيقة في استخراج الملامح منها لغرض تهيئتها لعملية التصنيف باستخدام Support Vector Machine (SVM) اثبتت التجارب ان النظام المقترح حقق أفضل النتائج في الصحة التي كانت 96.8% باستخدام polynomial kernel.

1. Introduction

Recently, there are great need for Document Image Retrieval (DIR) system because of the wide spread of electronic devices that facilitate the acquisition and archiving of documents, produces growing numbers of paper documents that are converted into electronic form. DIR system aims to find relevant documents from a corpus of digitized pages depending on the image features only. Generally, the user submits to the system a query and the result will obtained as a list of documents that match the query in specific features [1]. Many approaches have been proposed for searching and retrieving from document image collection that based on word spotting techniques. S. N. Srihari et al. [2] propose a system for spotting words in scanned in three document image scripts, Devanagari, Arabic and Latin. The query words that are searched in the document images are retrieved and ranked, where the ranking

*Email: Haithem_72@yahoo.com

measure is a similarity score between the query words and the candidate words based on shape features of spotting words. The performance of this proposed system is seen to be better for printed scripts as compared to handwritten text. T. Sari and A. Kefali [3] present a search engine for Arabic documents by proposing a method for indexing and searching degraded document images without recognizing the textual patterns. The proposed approach deal with textual-dominant documents either handwritten or printed. The proposed system was tested on some Arabic historical documents with recall of about 56.62% and a precision approximating 77.78%. F.Zirari et al. [4] propose a methodology to spot words in historical Arabic documents. Elastic Dynamic Time Warping was used for matching the shape features between words. The proposed words spotting system achieved recall rates 95.75% on average while keeping the precision at 96.47% on average. M. Khayyat et al. [5] develop a learning-based word spotting system for Arabic handwritten documents by adapting the nature of Arabic handwriting, which can have variable boundaries between words and sub words. This technique has performance a recall of 96.0% for recall and 95.4 for precision.

2. Problem Statement

Recently the amount of printed Arabic documents exist in electronic form are growing rapidly. This has created an increasing demand for adaptive retrieval methods to relevant information. In this paper, a printed Arabic documents retrieval system has been proposed that based on extracting significant features from header-words.

3. The Proposed Work

The proposed Arabic document retrieval system has several major steps. These steps are header-words segmentation, preprocessing, feature extraction, classification and retrieval. The input to the system is a printed Arabic document where its header words are extracting by segmentation process then pass into preprocessing stage. After that, the system extract features for each extracted header-words and classify these them based on their class labels. The last step of the system is to retrieve the desired documents based on the class lable. The overall system steps are illustrated in Figure-1. For the proposed retrieval system, a dataset of printed Arabic document images has been constructed.

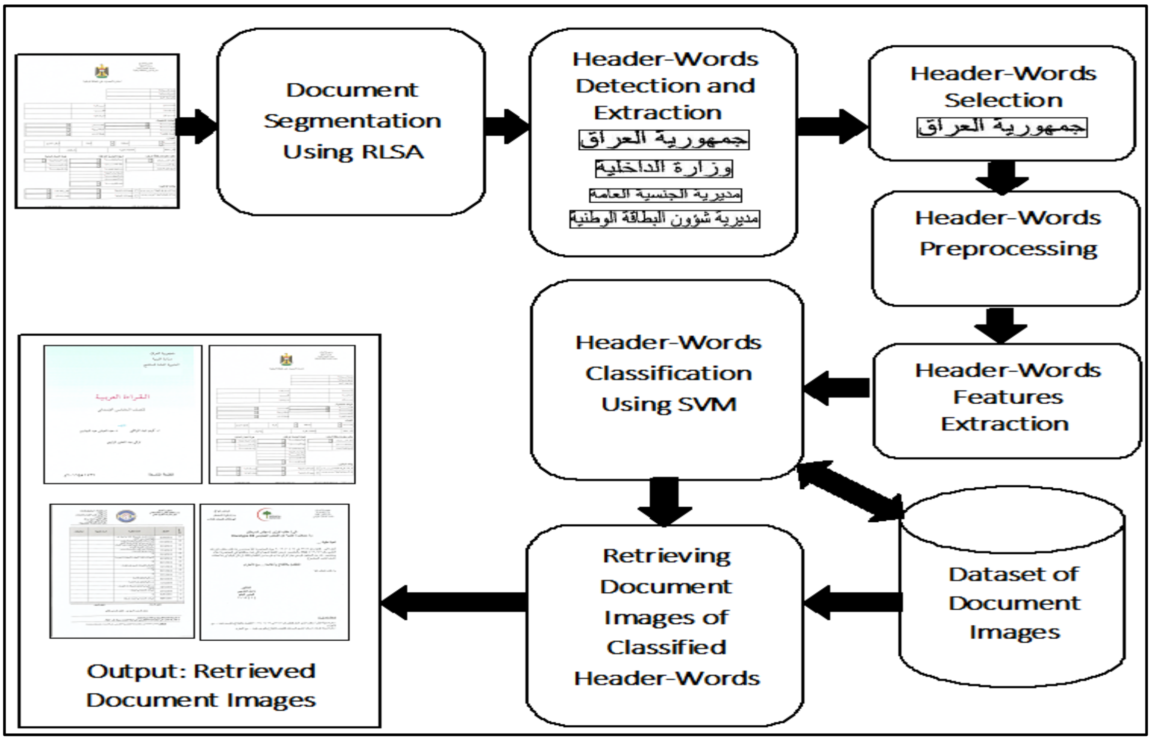

Document Segmentation Using RLSA
Header-Words Detection and Extraction
جمهورية العراق
وزارة الداخلية
مديرية الجنسية العامة
مديرية شؤون البطاقة الوطنية
Header-Words Selection
جمهورية العراق
Header-Words Preprocessing
Header-Words Features Extraction
Header-Words Classification Using SVM
Dataset of Document Images
Retrieving Document Images of Classified Header-Words
Output: Retrieved Document Images

Figure 1- Main steps of the proposed Arabic documents retrieval system.

3.1 Dataset

An official printed Arabic document dataset has been constructed and tested in the proposed retrieval 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 store 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-1 shows samples of Arabic document images dataset of different categories.

Table 1- Samples of document images in the dataset.

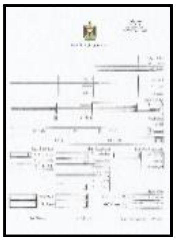
Class	**Sample**	**Class**	**Sample**	**Class**	**sample**
Report		Form		Letter	
Book		Announcement		instructions	

Table- 2 shows the classification of constructed Arabic documents dataset according to their color and resolutions.

Table 2- Samples 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
Instructions	16	6	14	4	8	4	52
Total	86	34	72	24	63	23	302

3.2. Segmentation

For sentence level segmentation, Run Length Smearing Algorithm (RLSA) has been modified and applied to segment document images [6]. 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 to estimate the value of smeared threshold value. 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.

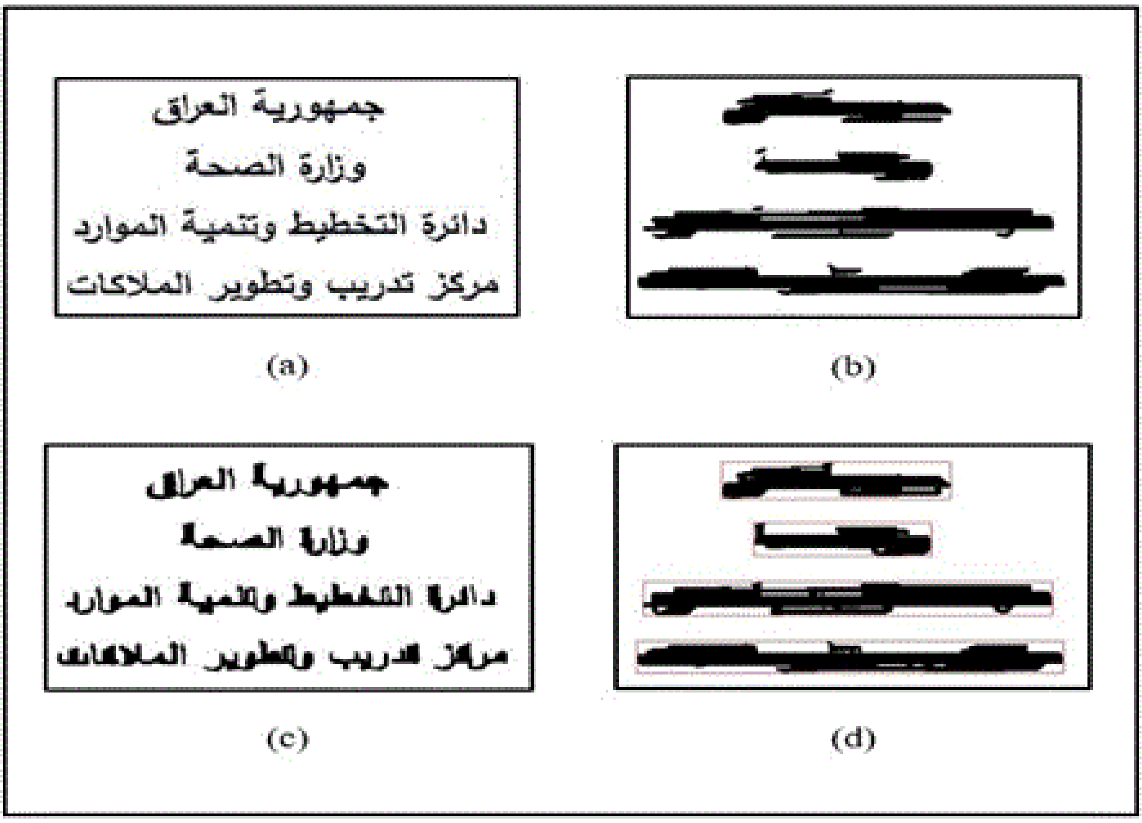


Figure 2-Document segmentation using RLSA technique. (a) A part of binary document image, (b) Image after horizontal smearing, (c) Image after vertical smearing, (d) Image after logical AND between horizontal and vertical smeared images.

3.3. Preprocessing

Preprocessing is an essential stage of any retrieval system that can be performed after the acquisition process. The preprocessing is designed to prepare the image of the route to the next stage of analysis. It is essentially to reduce the noise superimposed data and keep as much as possible significant information as presented. Generally, the preprocessing operations used include image normalization, image binarization, noise removal, and image thinning.

3.3.1. Image Normalization

The printed Arabic images in the used dataset have various sizes and resolutions. The retrieval systems are sensitive to small variations in the size and position as is the case in matching templates and correlation methods. Normalization of images size seeks to reduce variations between images due to the size of Arabic sentence to improve the performance of the classifier. Therefore, in the proposed system all the images are normalized into size 64 of the height and make the width flexible according to the image content to preserve the aspect ratio of the sentence shape. An example of the normalization process is shown in Figure- 3.

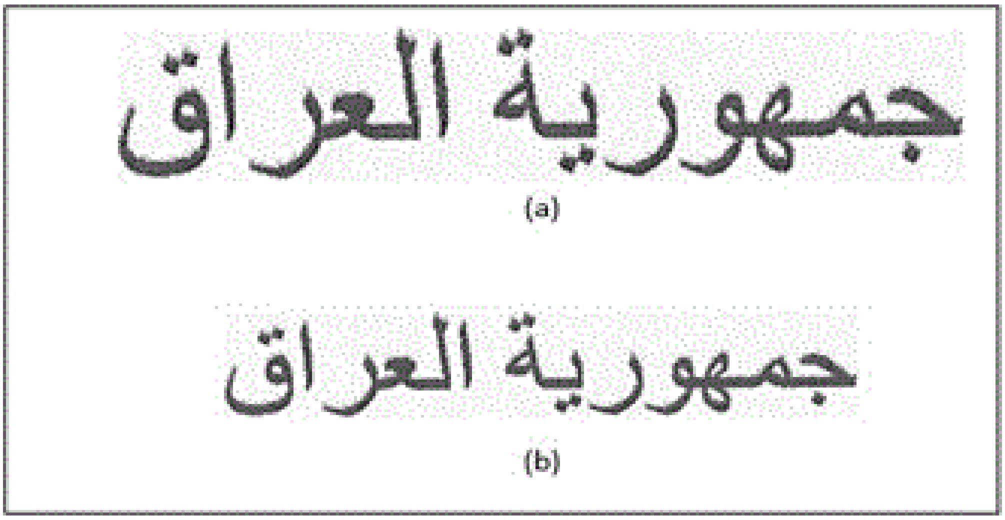


Figure 3- Image normalization. (a) Original image, (b) Normalized image.

3.3.2 Image Binarization

Binarization is the process of converting the gray image into a binary that is composed of two values 0 and 1, which make the image easiest to process. In general, using a binarization threshold appropriate reflecting the limits of high and low contrast in the image. In the proposed system, Otsu method [7] is used to convert the input Arabic image into binary. In addition, Median filter is used to remove unwanted pixels from the binary image. The result of applying the binarization process illustrated in Figure- 4.

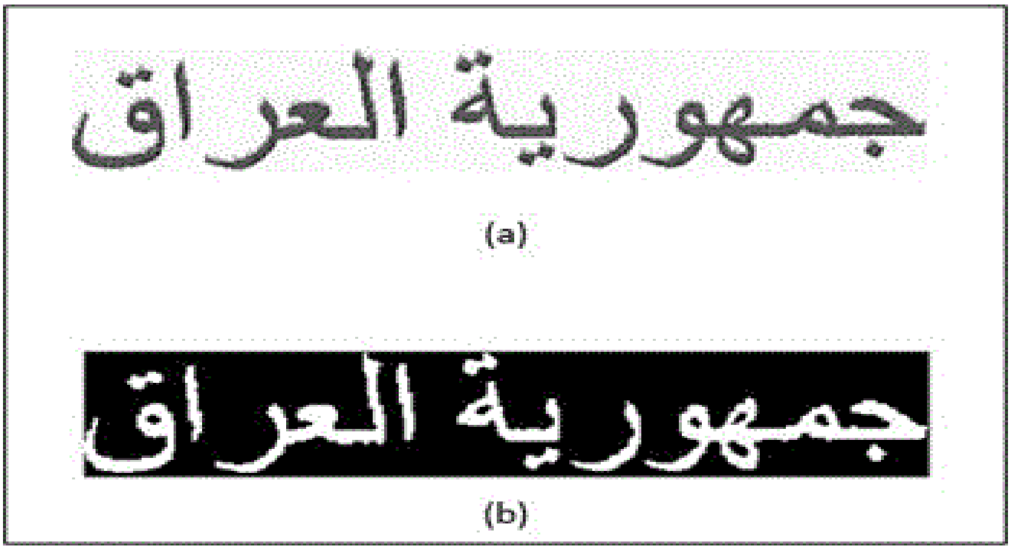


Figure 4- Image binarization. (a) Original image, (b) Binary image.

3.3.3 Image Thinning

Thinning algorithm is a morphological operation that is used to remove selected foreground pixels from binary image. It preserves the topology (extent and connectivity) of the original region while throwing away most of the original foreground pixels. The most common thinning algorithms are Stentiford, Zhang-Suen and Guo-Hall algorithms [8]. In the proposed system Zhang-Suen thinning algorithm is used and the result of applying this algorithm shown in Figure- 5.

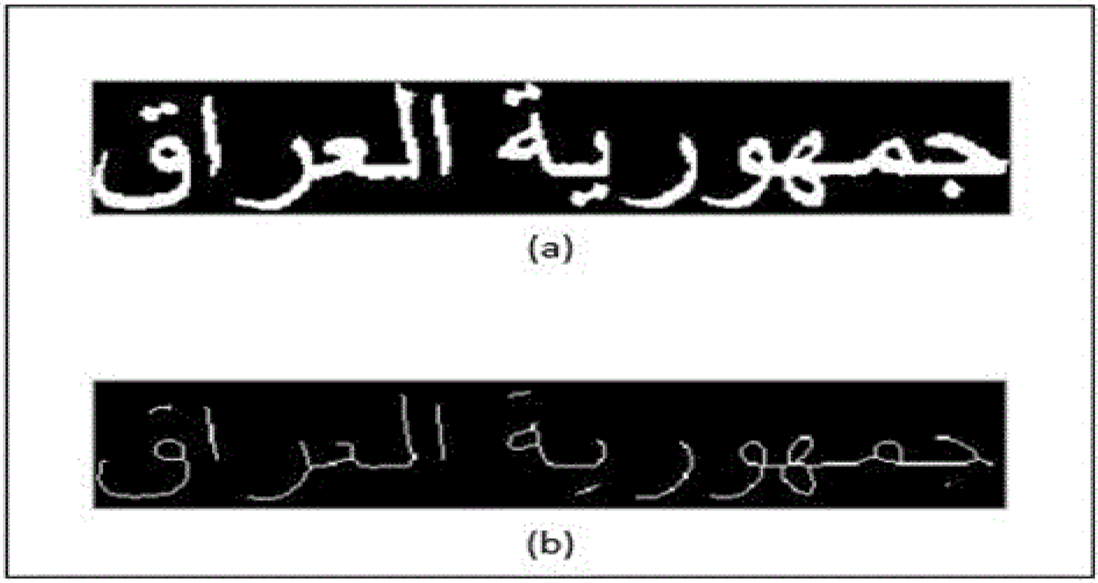


Figure 5- Image thinning. (a) Original image (b) Thinned image.

3.4 Features Extraction

n order to extract appropriate features that represent the input Arabic header -words, an accurate features extraction algorithm has been proposed. The proposed algorithm extracts the features based on the gradient image through several steps. First, the thinned image that obtained from the previous stages is used to find the image edges. In the proposed work, Roberts's edge detector [9] is used as in equations 1 and 2, and the resultant images of applying the filter are shown in Figure- 6.

$$G_x=\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix} \tag{1}$$

$$G_y=\begin{bmatrix}0 & 1\\-1 & 0\end{bmatrix} \tag{2}$$

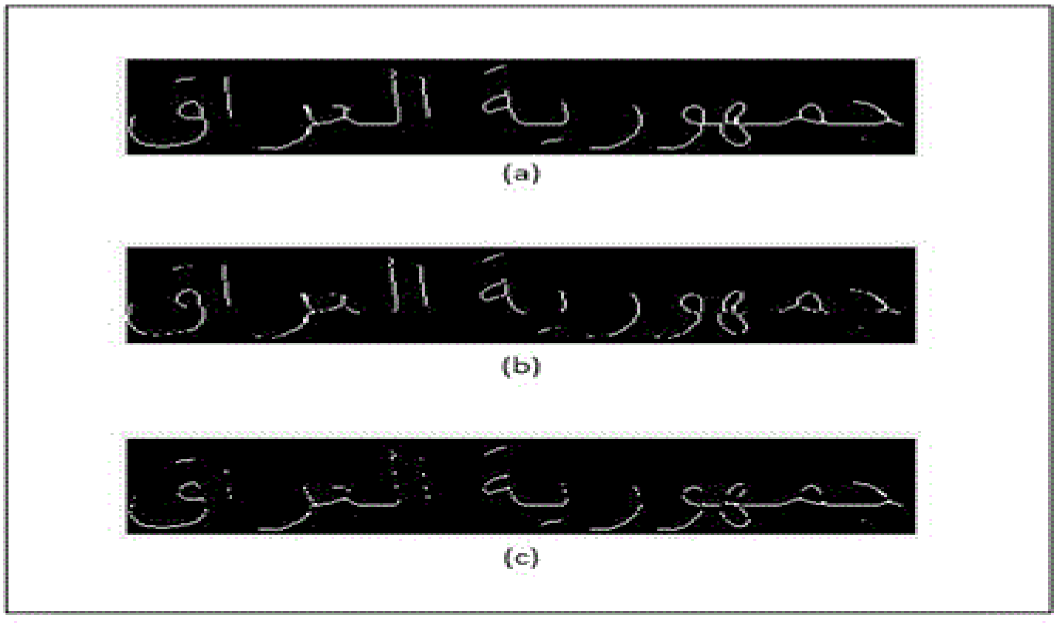


Figure 6- Edge detection. (a) Original image, (b) X-axis, (c) Y-axis.

The next step is computing the image magnitude and orientation using equations 3 and 4. The gradient magnitude and direction are shown in Figure- 7. Where g is gradient magnitude and θ is gradient orientation.

$$g = \sqrt{g_x^2 + g_y^2} \qquad (3)$$

$$\theta = tan^{-1}\left(\frac{g_x}{g_y}\right) \qquad (4)$$

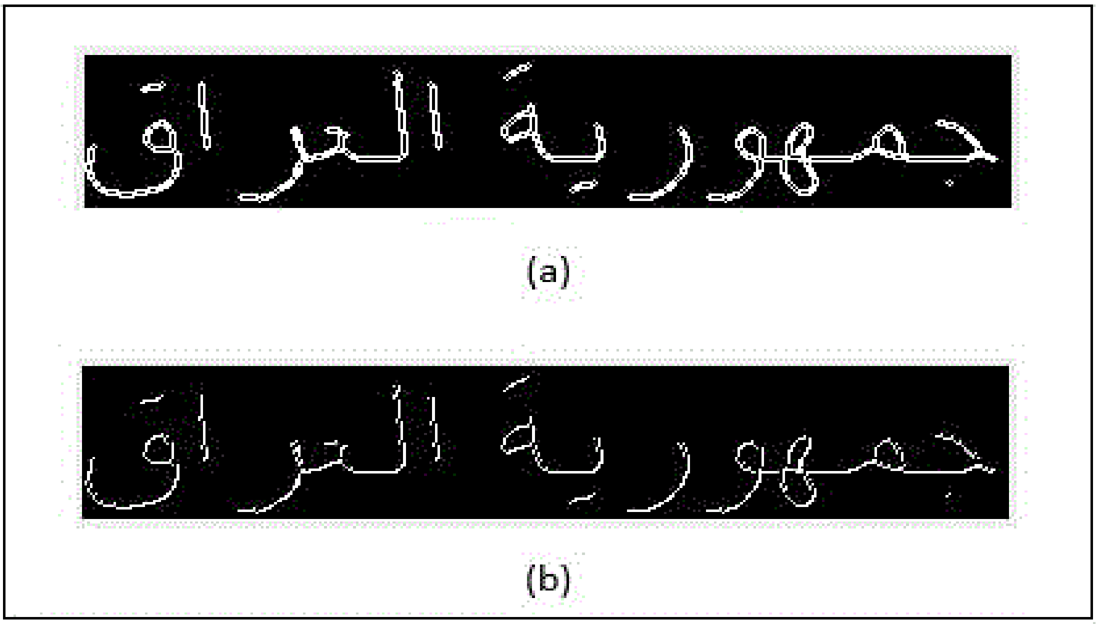


Figure 7- Image gradient. (a) Image magnitude, (b) Image direction.

Moreover, the gradient image is divided into four equal blocks as in Figure-8.

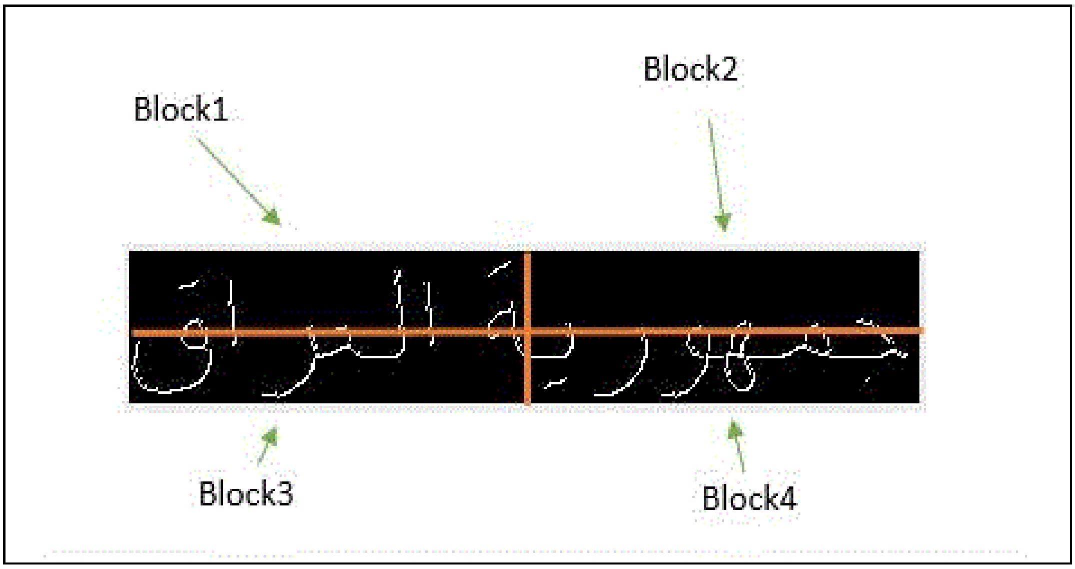


Figure 8- Blocks division.

For each block, the histogram of the orientation gradient is obtained. Since the output range of the gradient orientation fall in [-π — π] which gives many orientations. In the proposed system the gradient orientation is quantized into 5 orientation which are [0 30 45 60 90] as shown in Figure-9.

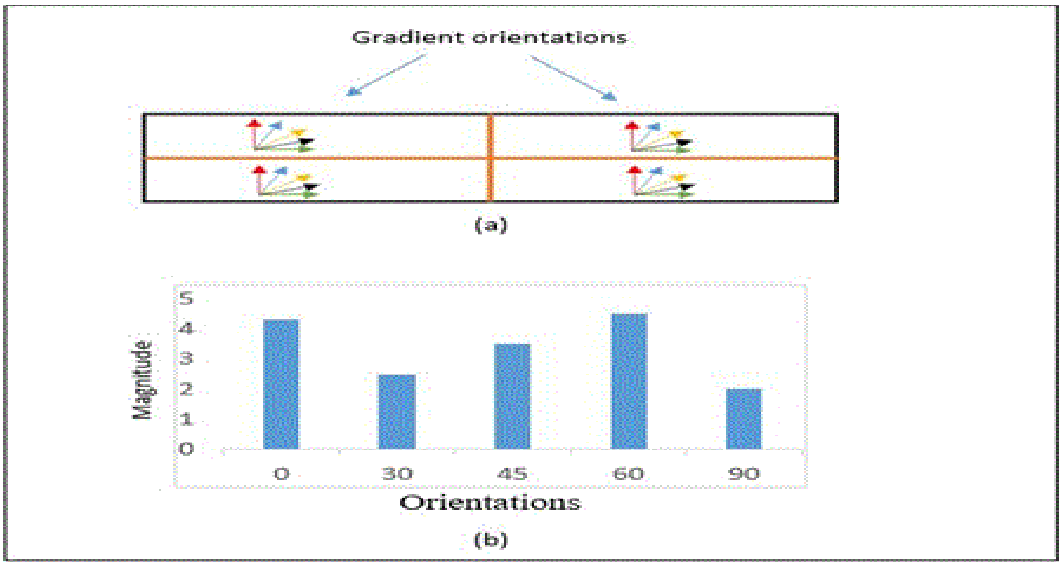


Figure 9- Gradients orientation. (a) Gradient orientations of four blocks, (b) Oriented gradients for one block.

In addition, the output of all blocks are concatenated to one dimension vector that represent the image features. The main steps of the proposed features extraction are shown in Algorithm 1.

```
Algorithm 1: Proposed Features
Input: Thinned Image
Output: Features Vector

Step1: Read the input image (I)
Step2: Detect the image edges by Roberts filter
Step3: Compute the gradient magnitude $I_m$ and orientation $I_o$ (Equations 3 and 4)
Step4: Divide the gradient image into 4 blocks
Step5: Scan the image and find the angles [$0^o$ $30^o$ $45^o$ $60^o$ $90^o$] for each block
Step6: Compute the weight of the chosen angles in each block
Step9: Combine the obtained weights into one dimension array (1*20)
Step8: Return (Features Vector)
```

3.5 Classification Using SVM

Classification plays an important role by assigning an example to an unknown predefined class from the description in the form of parameters. In the proposed system, Support Vector Machine (SVM) is used [10]. SVM is a binary classification method for supervised learning that was introduced by Vapnik in 1995. This method is therefore a recent alternative for classification. It is based on the existence of a linear classifier in a suitable space. Since it is a classification problem with two classes, this method uses a training set to learn the model parameters. It is based on the use of so-called kernel functions (kernel) which allows an optimal data separation. SVM is particularly effective in that it can deals with problems involving large numbers of descriptors, provides a unique solution (no local minimum problems like neural networks) and provided good results on real problems. In this paper one-vs-all multi-class SVM approach is used.

3.6 Retrieving Document Images.

The last step of the proposed system is retrieving the desired Arabic documents. The output of SVM is a class label that represent one of the Arabic header-words. Thus, all the document images that belongs to the predicted class are retrieved from the used dataset.

4. Experimental Results and Discussions

The proposed AADIR system was implemented in Matlab 2016a. The experiments are tested on an Intel Core i7, 64-bit Operating System, 2.50 GHz processor and 12 GB RAM. However, in the proposed system different kernels of SVM are used. SVM is commonly used with linear, polynomial and RBF kernels. Table- 3 shows the accuracy rate of applying the three kernels of SVM classifier. A multiclass SVM classification has been used in the proposed system and the experimental results demonstrate that it achieved a very high classification accuracy using the polynomial kernel.

Table 3-Comparison of retrieval results of different SVM kernels.

SVM Kernels	Accuracy
Polynomial	**96.8%**
RBF	93%
Linear	92.2%

5. Conclusions

In this paper, a printed Arabic document images retrieval system was proposed. The proposed system implement a new approach to retrieve the desired documents by extracting significant features from the header-words of document images. Furthermore, the retrieval system was tested by constructing a dataset of official printed Arabic document within 70% for training and 30% for testing in order to evaluate the system process. In addition, the performance evaluation of features extraction method prove that the retrieval system achieved very satisfied results of retrieving documents using SVM classifier with polynomial kernel.

References

1. Giotis, A. P., Sfikas, G., Gatos, B., and Nikou, C. **2017**. A Survey of Document Image Word Spotting Techniques. *Pattern Recognition*, **68**: 310-332.

2. Srihari, S. N., Srinivasan, H., Huang, C., and Shetty, S. **2006**. Spotting Words in Latin, Devanagari and Arabic Scripts. *Indian Journal of Artificial Intelligence*, **16**(3): 2-9.

3. Sari, T., and Kefali, A. **2008**. A Search Engine for Arabic Documents. In Colloque International Francophone sur l'Ecrit et le Document (CIFED): 97-102.

4. Zirari, F., Ennaji, A., Nicolas, S., and Mammass, D. **2013**. A Methodology to Spot Words in Historical Arabic Documents. In Computer Systems and Applications (AICCSA), In Computer Systems and Applications (AICCSA), ACS International Conference on IEEE: 1-4.

5. Khayyat, M., Lam, L. and Suen, C. Y. **2014**. Learning-Based Word Spotting System for Arabic Handwritten Documents. *Pattern Recognition*, **47**(3): 1021-1030.

6. Rege, P. P. and Chandrakar, C. A. **2012**. Text-Image Separation in Document Images Using Boundary / Perimeter Detection. *ACEEE International Journal on Signal and Image Processing*, **3**(1): 10-14.

7. Sehgal, S., Kumar, S. and Bindu, M. H. **2017**. Remotely Sensed Image Thresholding Using OTSU & Differential Evolution Approach. In Cloud Computing, Data Science & Engineering-Confluence, 7th International Conference on IEEE: 138-142.

8. Gramblicka, M. and Vasky, J. **2016**. Comparison of Thinning Algorithms for Vectorization of Engineering Drawings. *Journal of Theoretical and Applied Information Technology*, **94**(2): 265-275.

9. Liu C., Nakashima K., Sako H. and Fujisawa H. **2004**. Handwritten Digit Recognition: Investigation of Normalization and Feature Extraction Techniques. *Pattern Recognition*, **37**(2): 265-279.

10. Pasolli, E., Melgani, F., Tuia, D., Pacifici, F. and Emery, W. J. **2014**. SVM Active Learning Approach for Image Classification Using Spatial Information. *IEEE Transactions on Geoscience and Remote Sensing*, **52**(4): 2217-2233.