IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 21, 2021, accepted March 4, 2021, date of publication March 17, 2021, date of current version April 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3066477

Arabic Documents Information Retrieval for Printed, Handwritten, and Calligraphy Image

HASSANIN M. AL-BARHAMTOSHY[1], (Fellow, IEEE), KAMAL M. JAMBI[2], SHERIF M. ABDOU[3], AND MOHSEN A. RASHWAN[4]

[1]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[3]Information Technology Department, Faculty of Artificial Intelligence, University of Cairo, Giza 12613, Egypt
[4]Electronics and Communication Department, Faculty of Engineering, Cairo University, Giza 12613, Egypt

Corresponding author: Hassanin M. Al-Barhamtoshy (hassanin@kau.edu.sa)

This work was supported by the National Plan for Science, Technology and Innovation (MAARIFAH) – King Abdulaziz City for Science and Technology -the Kingdom of Saudi Arabia under Award 11-INF-1997-03.

ABSTRACT This paper presents a new computational backend model that supports Arabic document information retrieval (ADIR) as a dataset and OCR services. Therefore, different services that support document analysis, retrieving, processing including dataset preparation, and recognition will be discussed. Consequently, ADIR services provide general functions of the Arabic OCR to compose many other services in the OCR domain. Furthermore, the proposed work can provide accessing different methods of document layout analysis with a platform where they can share and handle such methods (services) without any setup requirements. One of the used datasets composed from 16,800 Arabic letters written by 60 writers. Each writer wrote each letter from Alif to Ya 10 times in two forms. The forms were scanned at 300 DPI resolution and are segmented in two sets: training set with 13,440 letters for 48 images per class label, and testing set with 3,360 letters to 120 images per class label Convolutional neural network (CNN) is used and adapted for Arabic handwritten letters classification. In an experimental test, we showed that our results outperform 100% classification accuracy rate on testing images. Therefore, the ADIR services provide a ''service description'', which includes an interface and a server's URL. The interface allows communication process between clients and services. Although, in this article we evaluate IR results and compared them with respect to corrected equivalent.

INDEX TERMS Layout analysis, image processing, OCR, information retrieval (IR), segmentation, recognition, features extraction.

I. INTRODUCTION

Arabic language is the primary language in the Middle East, and therefore, it includes a large volume of data posted in information applications. The Arabic language is the most growing language over the web. Arabic document information retrieval (ADIR) has a very broad range of heterogeneous data with related analyzed methods. This work is a new research domain and it includes more challenges covered in [1], [2].

Many of repositories aims to documents exploitation and documents analysis in the digital age [3]. Therefore, they are needing to access datasets and methods using web services (like Simple Object Access Protocol (SOAP)) [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

Library digitization needs extra OCRing works to achieve high recognition accuracy, especially for Arabic documents archiving. However, the performance of the OCR systems depends on the number of fonts [4], [5]. Training of a single font is time consuming approach works, especially with the unknown fonts and historical documents. Document layout analysis in Arabic historical documents is still under investigation, due to the lack of transcribed data [5].

The paper will be organized as the following description. Section 2 introduces the literature review with related works for characteristics of Arabic in document analysis and information retrieval for Arabic manuscripts. Section 3 explains the layout analysis framework and describes the different design stages of the entirely proposed model. The classification model architecture with the documented models' algorithms will be illustrated in section 4. Section 5 describes the

Arabic image retrieving and evaluates the output results with the experimental indicators for Arabic documents retrieving. Conclusion and future work will be presented in section 6.

II. RELATED WORKS

OCR services gained many attentions in document image analysis in the digital age. The OCR service can be a web application that provides electronic services between two applications and/or systems.

Lamiroy introduced a framework for open data and document annotation to serve analysis and exploitation [4]. This framework provides hosting, distribution, reproducing data, and data annotations. Therefore, datasets and services' methods are important to be accessed through web services mechanisms or protocols like (SOAP).

So, the meaning of document imaging itemized using the following [6]: (1) document sharing, (2) document searching & indexing, (3) frequent document information retrieval, and (4) special compliance requirements.

There are subsequent tasks in the preprocessing phase, before the layout analysis and IR take place. These tasks include: binarization, noise removal, skew correction, page and zone segmentation [7]–[9]. Therefore, the following characteristics and challenges should be considered.

1- Connected letters and cursives styles: European languages can be written in separate letters (which are predominant in simplicity and printing for simplicity) or in continuous letters (often used in handwriting), whereas the languages of major nations in Far East Asia (such as Chinese, Japanese, and Korean) are always enclosed by a separate letter, as for Arabic calligraphy (and it is shared by Bengali and Hindi) it is always written in the continuous connection and in cursives styles [20].

From the point of view of any computer mechanism for identifying graphical patterns, it is with the installation of all other conditions that it is easier to recognize patterns which are separate from recognizing them or they are related to each other. Wherever in the case of communication, the issue of boundaries setting of each graphical symbol ≪Grapheme≫ must be solved, which is what researchers in this field call it ''segmentation''. In addition to the issue of ''recognition'' on the letter that each graphite symbolizes should be segmented separately. As can be traced from Figure 1 below, correct identification of symbols requires the proper identification of their boundaries but assigning these limits in turn requires knowing the symbols first!

Thus, it is inevitable to solve both issues of ''segmentation'' and ''recognition'', which doubles the challenge. This cannot be done with irregular lines such as decorative shown in figure 2.

2. Overlap between the boundaries of the graphemes: This increases the challenge presented in the previous point that we sometimes see some slight overlap between the boundaries of the grapheme. Some of the regular Arabic font is represented by figure 3 below.

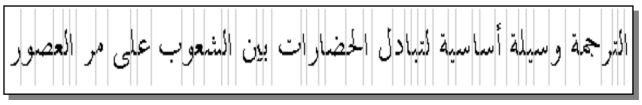


FIGURE 1. Horizontal communication between consecutive letters in the Naskh lines and delineation of its boundaries.

FIGURE 2. A decorative sample from the Diwani font.

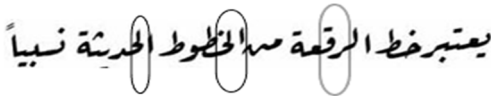


FIGURE 3. Overlap between the boundaries of the Arabic graphemes.

FIGURE 4. Changing Arabic character's drawing relative to its position.

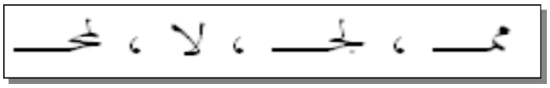


FIGURE 5. Examples of compound graphemes of more than one letter.

3 Changing the character's drawing with the change of its position in the word (Fig 4): This change is of course the result of writing related to fonts' letters, and this leads to a significant increase in the number of graphical symbols that must be dealt by any system for automatic recognition of written Arabic text compared to Latin language. European languages are printed using separate letters. From the point of view of computation mechanism to identify grapheme patterns, with all other conditions proven, the recognition accuracy is satisfied within fewer numbers of different graphic patterns and vice versa.

4 Compound graphemes of more than one letter: many of the transcription fonts that are used extensively in the competitions and computation of Arabic writing are needed. It contains two or three letter graphemes that can be handled as a unit. Figure 5 shows examples of some of these compound graphs. In addition to what we have mentioned in the previous point, these complexity in graphemes raise the number of graphical patterns that any system for recognizing the written text must deal with, which increase the degree of difficulty.

5. Arabic dotted: The greater the formal differences between graphical patterns, the more powerful all systems will be identified with all other conditions. And since a large percentage of the graphs of Arabic calligraphy are morphologically very similar and are distinguished only by the

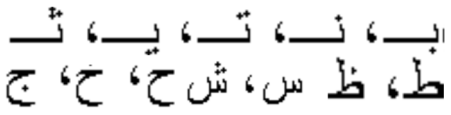


FIGURE 6. Examples of Arabic letters that are distinguished by dots.

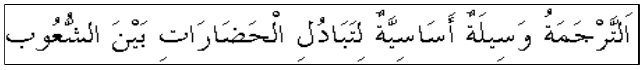


FIGURE 7. An example of an Arabic phrase with full sound tags.

presence or absence of dots. This certainly raises the challenge in front of any system to recognize the written Arabic text. Figure 6 shows some of the Arabic shapes with and without dots (below or over the letter).

6. Signs of phoneme (diacritical marks): diacritics in Arabic script are an additional complication in front of any OCRing system for recognition of Arabic text, because they do not fall into the context of a horizontal sequence such as alphabet [20] (they are in vertical positions above or below them) as shown in figure 7 below. Therefore, such OCRing systems deal with diacritics either by trying to discover them early and then delete them before the process of recognition or ignoring them on the grounds that contemporary Arabic writing rarely adds to these signs except for educational purposes or when quoting religious or heritage texts.

7. Text and non-text regions: The layout idea is the arrangement of the image or document in a readable way and understandable format. An Arabic document (Al-Watan and Asharq Awsat news)[1] are presented in Fig 8.

Accordingly, journals, news, early printed, machine printed, handwritten and ancient documents can be skewed, noisy and sometimes overlapped regions. It is also needing to separate each region from the others, and to deal with the structure of the page. Further, repeated data of the header and footer for a specific documents' categories need extra processing [9], [10].

Although, Arabic historical images are very important in human wealth and intellectual production sources [20]. The article presented a novel dataset based on historical documents and homogeneous and geometric layouts.

It is important to remind the segmentation and recognition of the OCRing are our aim, and text processing is not our aim of this paper. However, graphics processing will be handled and briefly explained.

III. ARABIC DOCUMENTS INFORMATION RETRIEVAL FRAMEWORK

The layout analysis based on document image descriptor in features vectors or quantifiers descriptors is described. Many preprocessing tasks and related models are needed to analyze documents in the following description. The pre-processing phase will be described in the following algorithm.

[1] http://www.alwatan.com.sa/Default.aspx? and https://aawsat.com/

Many of methods are used in document layout analysis; the first method is to differentiate between text and non-text objects in the image using segmentation module. The segmentation module uses algorithm(s) to differentiate between the text and non-text objects. The segmentation process of the text object is employed by surrounding the text objects by boxing or polygons. In addition, this can be done by word, line and paragraph objects. Consequently, an appropriate kernel with inline functions are used with trial and error approaches using the following ADIR algorithm:

Step 1. Assign the input and output path to read and store the scanned documents.
Step 2. Read the input document or the scanned image.
Step 3. Binarize and convert the input images into grayscale images (using binarization algorithm)
  Step 3.1. Clean the input images by using suitable methods (e.g., Otsu's)
  Step 3.2. Check the skewing (and de-skewing if needed) angles of the input images (or regions).
Step 4. Assign the kernel according to the segmentation method (word, line, and paragraph segments)
Step 5. Use dilation, erosion closing and opening morphological operation to modify the images (if needed).
Step 6. Find contours and draw colored box around each object.
Step 7. Store the output results into output files.

Methods of automatic recognition of Arabic manuscripts (printed and handwritten) can be classified into the following.

1- Segmentation methods for the Arabic manuscripts (printed/handwritten texts).
2- Methods of recognizing the handwritten Arabic writing.
3- Building linguistic resources to train and evaluate Arabic writing recognition systems.

Automated text recognition framework belongs to a broader field of applied in ''Pattern Recognition'', and the functional architecture of these systems can be placed in the general framework. Where ''analog signals'' (corresponding to the patterns to be recognized), which are the printed/written texts in our case, are converted to ''Digital Signals'' and then provided with the computer, then ''Preprocessing'' for these digital signals is performed, such as the exclusion of some formats. A fold of noisy patterns of these signs are then extracted (i.e. a set of mathematical characteristics that are unique to them). After that, the features extraction module takes place.

It is characterized by the training path where mathematical models (often statistical) are being constructed from the features extraction of the signals corresponding to the patterns of training samples, and then these models are saved efficiently in a database (dataset) to be called in one of the classification mechanisms that decide which patterns are closest to the signals corresponding to the input patterns to be recognized.

This framework reflects the ''machine learning'' theory (which is applied as appropriate to approach issues that

FIGURE 8. Arabic Documents: Samples showing typical problem to separate text, non-text and possible noise regions.

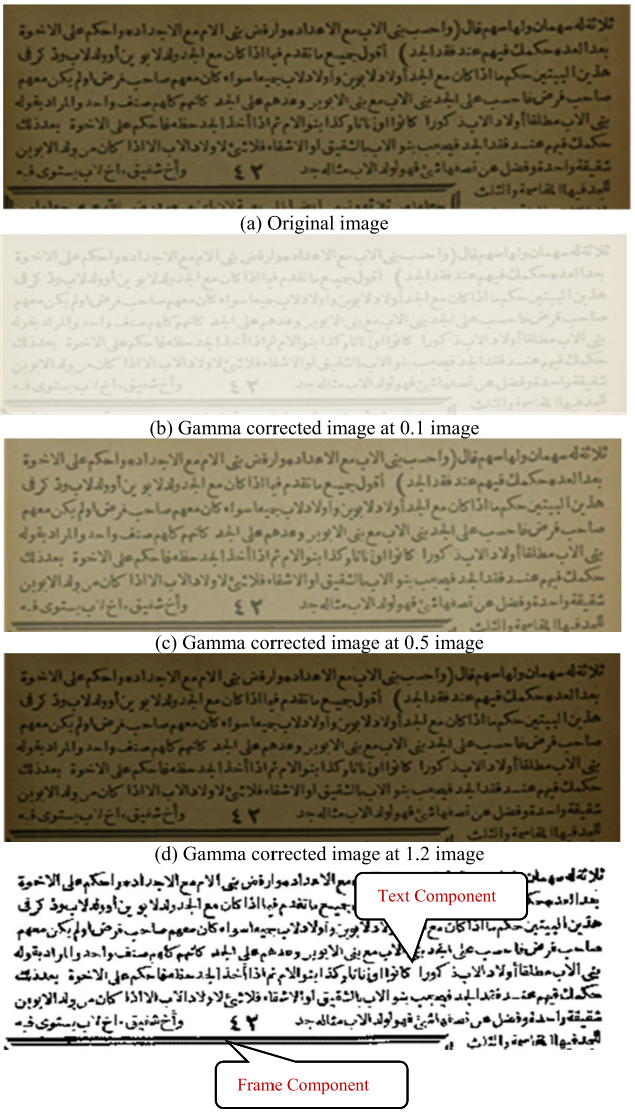
are not known to it or where it cannot be obtained by ''Closed-Form Solutions''). Various computer learning methods emerge from the principle of the ability to learn by repeating exposure from both the correct examples and wrong examples or by repeating the exposure to questions and their answers about the particles of the issue to be approached. For example, a child may be able to read the text written in his mother tongue before he learns the foundations and rules of the language through imitating the older and through attempts of right and wrong.

In general, the activation of these methods is based mathematically and computationally for this principle on extrapolating the probabilistic context of words and their letters (corresponding to the patterns we study in this paper). Rather than their linguistic context to reach the calculation of the mathematical probability of the occurrence of every possible identification of the word between what precedes it and what follows it likelihood of identification with the highest probability. This calculation requires the formation of a probability model that simulates all the sequences of linguistic units as they occur in real-life language use.

Whatever mathematical and computational methods are chosen to construct such a probabilistic model, they must have empirical data that fill a wide range of patterns in parallel with the series of patterns codes corresponding to these patterns as directed computer learning requires and is called a process. Running these mathematical methods on the contents of this container to build the probabilistic model is called ''Training.'' The contents of this container are also called a ''language resource''. Of course, the probabilistic model will carry the statistical properties of the resource, and which it in turn to carry a statistical and contextual characteristic of patterns which are expressed as we are going to in the next section.

A. DOCUMENT IMAGE PREPROCESSING

Generally, the dataset includes also historical Arabic images collected from different national and significant libraries over the world. Therefore, ''gamma transformation or intensity transformation operations'' is used to eliminate these effects, in such way to improve image quality and to achieve light balance. Accordingly, Ma and et al [19] used equation 1 to



(a) Original image

(b) Gamma corrected image at 0.1 image

(c) Gamma corrected image at 0.5 image

(d) Gamma corrected image at 1.2 image

(e) Results of Gamma transformation after connected component

FIGURE 9. Example results for document image processing using Gamma Transformation.

calculate the document:

$$D_{out} = A.\mathbf{D}_{in}^{\gamma} \tag{1}$$

where γ represents the brightness of the document. When γ <1, the image after ''Gamma transformation'' becomes brighter, and when γ > 1, the image becomes darker. By adjusting γ according to the experiment, γ varies between 0.4 and 0.5, as shown in Figure 9 (a) Original, (b), At 0.1 (c) At 0.5, and (d) At 1.2. Finally, the connected component used 8 neighborhood seed filling tasks in our work to standardize the result of processing based on the following rules is shown in figure 9 (e).

B. TEXT LINE SEGMENTATION

Graph model is used for text-line segmentation in historical Arabic documents. Therefore, the shortest path is found in

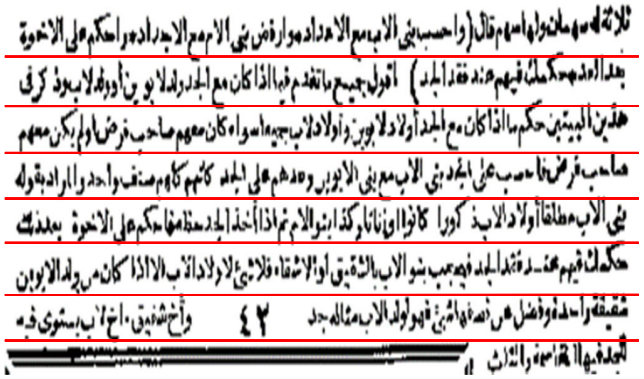


FIGURE 10. The segmentation graph of historical Arabic document image.

the graph model in our case of text line segmentation. So, the proposed method is defined in figure 10 and as the following steps.

Step 1. Let's suppose that the area of the document image is denoted as A, and let's α represents the threshold of frame (classified as connected component).

Step 2. If the width and height ratio or height and width of the threshold $> \alpha$, the component is classified as Frameline.

Step 3. Otherwise the component is classified as text.

C. EXPERIMENTAL ARABIC CHARACTERS DATASET FOR HANDWRITING

The best results have been achieved over the past two decades in dealing with this issue through mathematical equations according to one of two methodologies [1], [12]; Hidden Markov Models (HMM) and Artificial Neural Networks (ANN), which draw inspiration from the mechanisms of real neurons in terms of arousal and response and their connections together in dense networks. Where each of these artificial neural networks acts as a function that correlate between the Arabic writing process and the observations (observed because of the training process) - as mentioned in the previous section.

Many types of these artificial neural networks, each of which are appropriate for a specific range of issues, are known from those that cannot be obtained by ''Closed Form Solutions'' (Mathematical Laws), and these networks can be used to link in both directions between the inputs and the observations they lead to. Observed - that is, to get one of the two groups with the knowledge of the other group - and of course the reverse link is the tool used to recognize the letters of the writing corresponding to our numbered curves that represent the tracking of handwriting while writing. Large amounts of inputs in parallel with the corresponding observable observations is experimented.

The experimental dataset contains Arabic characters images that are normalized and converted into 16 × 8 pixels with related label features (name of the letter in the image). The proposed work of the Arabic character recognition transcribe character contained in the scanned images into digital characters. The partition method into training and testing was

TABLE 1. An experimental results of the dataset using classification using decision tree for machine learning.

Arabic Character	Letter Name in English	Precision	Recall	F_1
أ	Alif	74	71	73
ب	Bā	69	64	66
ت	Tā	76	77	76
ث	Thā	34	60	44
ج	Jīm	61	56	58
ح	Hā	79	80	83
خ	Khā	60	55	59
د	Dāl	66	66	66
ذ	Dhāl	45	44	44
ر	Rā	76	76	76
ز	Zāy	55	58	54
س	Sīn	50	41	54
ش	Shīn	69	69	69
ص	Sād	73	68	70
ض	Dād	69	68	68
ط	Tā	26	27	27
ظ	Zā	42	30	35
ع	Ayn	57	56	56
غ	Ghayn	52	46	49
ف	Fā	48	53	50
ق	Qāf	65	65	65
ك	Kāf	66	64	65
ل	Lām	81	85	83
م	Mīm	82	83	81
ن	Nūn	82	82	82
ه	Hā	49	65	48
و	Wāw	86	84	85
ي	Yā	64	64	64

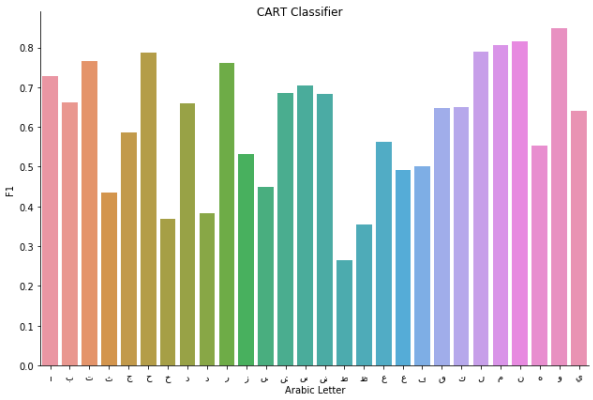


FIGURE 11. An Experimental Classification Results of the Proposed Dataset using Decision Tree for Machine Learning.

kept as standard for the proposed dataset (70% and 30% respectively).

These Arabic characters are segmented and extracted from many other corpora that includes images of words. Logistic regression (LR), Support Vector Machine (SVM), Naïve Bayes, and decision Trees are used as different classifiers to work with. Therefore, precision, recall and F_1 measures are computed as shown in Table 1, and illustrated in figure 11 to represent F1 measure.

Replace the linear classifier with a neural network. Let's take a few minutes to dig into our neural network and see what it has learned by accessing the weights_ attribute of

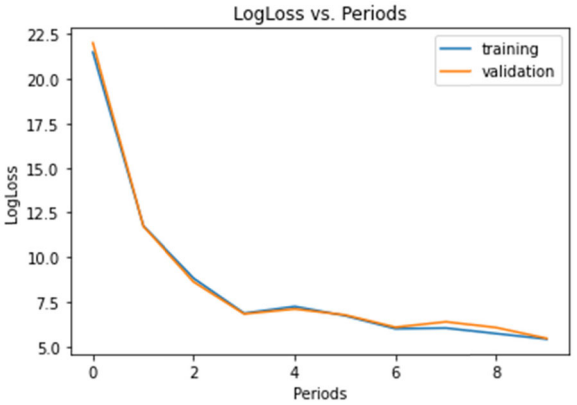


FIGURE 12. Decision Tree Classifier to train and validate Arabic Numerals.

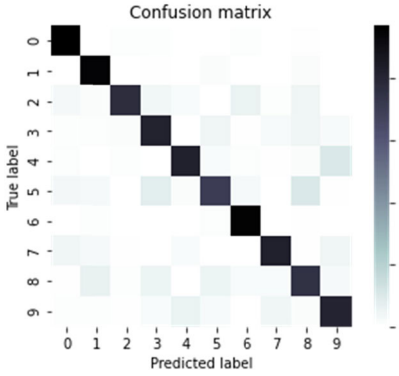


FIGURE 13. Confusion matrix of the experiment to classify the Arabic handwritten digits.

our model. The input layer of our model has 784 weights corresponding to the 28 × 28-pixel input images. The first hidden layer will have 784 × N weights where N is the number of nodes in that layer. We can turn those weights back into 28×28 images by reshaping each of the N 1×784 arrays of weights into N arrays of size 28 × 28.

In general, the activation of these methods is based mathematically and computationally for this principle on extrapolating the probabilistic context of words and their letters (corresponding to the patterns we study in this paper). Rather than their linguistic context to arrive at a calculation of the mathematical probability of the occurrence of every possible identification of the word between what precedes it and what follows it Likelihood of identification with the highest probability. This calculation requires the formation of a probability model that simulates all the sequences of linguistic units as they occur in real-life language use.

Whatever mathematical and computational methods are chosen to construct such a probabilistic model, they must have empirical data that fill a wide range of pattern fingerprints (questions) in parallel with the series of pattern codes corresponding to these fingerprints (their answers) as directed computer learning requires and is called a process. Running these mathematical methods on the contents of this container to build the probabilistic model is called "Training." The contents of this container are also called a "language resource." Of course, the probabilistic model will carry the statistical properties of the resource.

As a real testing, the classification report for three classifiers are done, as a real test figure 12 illustrates the Decision Tree Classifier to classify each handwritten of Arabic characters for precision, recall, and F_1 measure.

We are going to calculate the accuracy with a linear model of this handwritten forms. Figure 13 shows the results for this experiment with the hyperparameters for batch size = 10, learning rate = 0.02 and steps = 100 to classify the Arabic handwritten digits.

Fig 14 illustrates surrounding boxes for word, line, and paragraphs after ADIR processes of the proposed OCRing system. Regions of interest (text and non-text regions) are the important point to localize in document layout analysis.

(a) Word Layout (b) Line/Row Layout

(c) Margin Layout (d) Paragraph Layout

FIGURE 14. Illustrates surrounding boxes for word, line and paragraphs after segmentation process.

Recently, there are web services (RESTful)[2] with many useful tools are available on GitHub.[3]

D. DOCUMENT MODEL

The proposed dataset in this work is retrieved, filtered, annotated, and applied using machine learning algorithms, over the collected data. Gathering the data from internet using API

[2]https://www.digitization.eu

[3]https://github.com/impactcentre

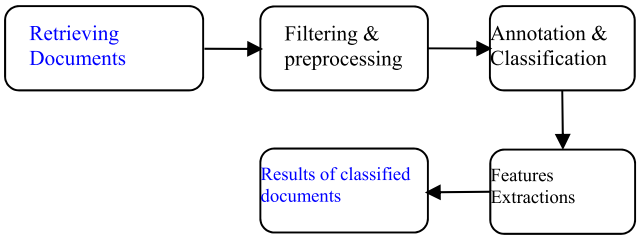


FIGURE 15. Visualizing dataset creation of the proposed work.

and creating layout-keyword related to Arabic documents is the first challenge. The second challenge is filtering and cleaning of the retrieved documents if the data has noises. The third challenge is annotating the cleaned documents in the dataset to their categories will be after filtering phase. After annotation phase, several machine learning algorithms can be applied to the annotated dataset using document features extractions. Fig 15 illustrates the dataset workflow of the proposed work.

The dataset is retrieved based on Arabic documents related to six domains, such as: news, scientific publications, early printed, printed, and handwritten. However, most of the retrieved documents were not expressed with any domain type. Accordingly, the domains of the documents can be involved to indicate the labeling of the documents. Table 1 describes the collected documents of the six documented domains, before and after the filtering and preprocessing phase.

The collected data contains noisy data, such noises data should be corrected or removed. The following items exemplified such noisy data.

1. Duplicate documents, which are retrieved more than once [11].
2. Retrieved documents that are not related to the predefined domains [12].
3. Spam documents that include advertisements or harmful links.

The objective of the document model is to analyze and understand the document structure (text and non-text regions) in order to extract the relevant information and create the internal representation of the layout. This representation is very important to understand the context meaning of the document. Consequently, the task of the document model includes layout structure: text and non-text regions. Therefore, the document model uses a well-defined extracted feature, and features vectors. The output results of the document model are used to achieve the domain type of the document (internal representation).

The proposed solution will be known the ADIR services, according to the following list of services:

- Dataset Binarization: updating images' colors to be images' grayscale into binaries images (black and white images).
- Noise Removing: Removing unnecessary, distorted data and special logos will be done before or after cleaning task [16].

TABLE 2. Collected arabic documented dataset and changes in the number before and after preprocessing.

Domain	Documents Filtering and Preprocessing	
	Before	*After*
News	100	90
Sientific Journals	100	95
Early Printed	150	100
Printed	200	180
Handwritten	100	75

- Data uploading: Uploading the binarized images or dataset into the hosting.
- Page Segmentation: Finding text and non-text regions, and segment text regions into segmented paragraphs, segmented lines, and segmented words.
- Features Extraction Model: Use training data to extract features and build language model to be used in recognition model.
- Recognition model: Use the trained features in the language model to recognize the text contents of the documents.

Posting a request for specific URL address. If data is available, documents will be provided by separate files. The binarization method (service) includes png and jpg into the ground-truth descriptions with the extension txt.

E. DATASET ANNOTATION AND TESTING

One of the used datasets composed from 16,800 Arabic letters written by 60 writers, 90 % of them using right-hand. Each writer wrote each letter from Alif to Ya 10 times in two forms. The form was scanned at 300 DPI resolution.

Convolutional neural network (CNN) is used and adapted for Arabic handwritten letters classification. The proposed CNN trained and tested using the whole 168,000 Arabic handwritten letters. The dataset is segmented in two sets: (1) Training set with 13,440 letters for 48 images per class label, and (2) Testing set with 3,360 letters to 120 images per class label. In an experimental test, we showed that our results outperform 100% classification accuracy rate on testing images.

The collected dataset has been annotated by three annotators by labelling each document to be one of the five classification domains. Some of these documents need to correct according. Table 3 illustrates the three experts (evaluators) about the collected documents' dataset. Therefore, the accuracy of each annotator can be calculated relative to each other. The three annotators asked to label the documents as News, Journal, Early-Printed, Printed, and Handwritten.

1. News agreement: the three annotators accepted 90 from100 images.
2. Journals agreement: 95 documents out of 100 agreed on a label of scientific journals.

TABLE 3. Opinion evaluation results of the collected dataset.

Collected Dataset Evaluation Results			
Annotators			*Final Decision*
First	*Second*	*Third*	
News	News	News	News
News	Journal	News	
Journal	News	Journal	Journal
Journal	Journal	Journal	
Early	Early	Early	Early
Eraly	Printed	Early	
Printed	Early	Printed	Printed
Printed	Printed	Printed	
Handwritten	Handwritten	Handwritten	Handwritten
Handwritten	Ancient	Handwritten	

3. Early-Printed: the three annotators accepted 100 documents from 150 images.
4. Printed: the three annotators found 180 documents out of 200 images.
5. Handwritten: Each annotator reviewed the 100 documented handwritten forms. They were encountered 75 documents.

A summary of the annotator's decision is shown in table 3.

F. DOCUMENT REGIONS CLASSIFICATION

This phase receives its input from document stream and splits the segmented regions into their component list of sub-regions (objects) or region of interests (RoI).

This task splits the Arabic document (corpus) into pieces called regions (words layout, line/row layout, page/margin layout, and paragraph layout). The splitting task of the segmented regions is performed by using specific threshold values (determined by a type of document). Therefore, many of Arabic documents are needed to create our corpus to work with segmentation and recognition (section D). Any proposed segmentation includes the following steps:

Step 1. While there is a stream for documents Do
  Step 1.1. Separate and split the input document stream out periods that exist at the end of each document.
  Step 1.2. Split and mark the object when find out any objects according to the determined threshold.
  Step 1.3. Split objects using standard layout method.
Step 2. Use the RoI to employ the object tagging (OT) for the entirely document.
Step 3. Recognize the object regions to be text or non-text regions.
Step 4. Stop.

Next section describes the ADIR services tasks to handle the server using the proposed interface. The only thing we need is to maximize the Arabic dataset to be very large. However, depending on how we plan to use our model, we need to be satisfied about the quality of the dataset we use. When in doubt, the general rule is the more data we have, is the better. Moreover, depending on the corpus size (documents content), training can take several hours or even days, but fortunately, we can store the analyzed data and extracted features on a storage disk. This way we do not have to do the analyzed tasks of model training every time we need to use it.

G. ADIR SERVICES

The proposed ADIR services providing, and interface allows document layout and OCRing operations to be combined with other services to provide new functionality of information retrieval. These services are shown in Fig 16.

Sometimes many of extraneous and unnecessary frames, lines and special graphics are not needed or not required, such as symbol tags and repeated data (such as header and footer). Therefore, such unnecessary data can be removed. Removing unnecessary and special logos will be done before or after cleaning task.

This task is important to avoid and recover the standard document of the Arabic. Some of contraction types can be classified into normal, negated, and colloquial. According to the Arabic document nature, the colloquial contraction with an acronyms or short end description (abbreviated). Other types such as euphemism taboo word (restricted), vulgar and accepted will be discussed in other literatures.

The management process of the proposed Arabic document information retrieval (ADIR) services is implemented according to the following scenario algorithm.

ADIR Services Scenario Algorithm

Step 1. While the user asks the ADIR service for information about a set of services (e.g. analyze, search, segment archive, and recognize) Do
  Step 1.1 The ADIR service collects the information for each service and sends it to the user, which chooses one of the following tasks on behalf of the user
    a. Refine the enquiry, or get more information, then repeat Step 1.1 again.
    b. Make the selected service
    c. exit
  Step 1.2 The user requests the selected service, and the ADIR service checks availability.
  Step 1.3 Either all are available, either alternatives are offered to the user, which go to back to step 1.2; or the user goes back to step 1.
  Step 1.4 Make and execute the selection
Step 2. Give the user a result of the execution as a confirmation.
Step 3. The user can select or modify or cancel his request.

IV. CLASSIFICATION MODEL ARCHITECTURE

A. DATASET AND METADATA DESCRIPTION

Different categories of the Arabic documents will be used, such as news pages, journals pages, early printed, printed,

handwritten, and ancient documents are considered. The dataset includes ideal pages, rotated or skewed pages, distorted pages, scaled pages and degraded pages.

The presented dataset may have XML-based formulation metadata as a part of the description. The metadata description includes the following properties:

- Bibliographic description: such as title, publication date, document type, page number, ...etc.
- Physical description: language, typeface, fonts, number of columns, ...etc.
- Digitization description: image resolution, scanner used, dimensions, bit depth, file type, source of digitization, ...etc.

B. ADIR AUTOMATION ARCHITECTURE

Usually, supervised, and unsupervised machine learning algorithms are relevant to classify such input documents. In addition, reinforcement and semi-supervised learning may be used.

To describe the classification process of documents in scientific notation. The Arabic documents (D) is a set of combined texts and labels. $D = \{(D_1, l_1), (D_2, l_{2)} \ldots, (D_n, l_n)\}$ where, $D_1, D_2, \ldots, D_n$, are a list of documents, and their paired labels are classes: $l_1, l_2, \ldots, l_n$. If the learning algorithm L is trained with the training dataset D, the classifier φ such that $F(D) = \varphi$ is used. The workflow of the proposed automated Arabic documents dataset classification in ADIR is illustrated in Fig 17.

Usually, the dataset is divided into two datasets, training, and testing datasets. The preprocessing phase and the features extraction phase are overlapped to be used in both training and testing. In the training phase, each document has its own equivalent type (category or domain class) that was labeled before. The cleaning documents will be passed to the feature extraction phase to extract significant features (numeric arrays or vectors).

These significant features (vectors) are feeding with the corresponding related labels to the machine learning algorithm to learn various documented patterns related to each category or domain and combine classification model. This learned knowledge will be used to predict categories for new documents. Once we have the classification model (working model), we can test such model using accuracy metrics.

C. FEATURE EXTRACTION (FEATURE ENGINEERING)

Feature are measurable properties for every data element in a dataset. The feature extraction (engineering) is the process to transform the input stream into a set of measurable value (numerical values). These features can be extracted using machine learning algorithm to differentiate and recognize between the documents' dataset.

''Vector Space Model'' or ''Term Vector Model'' is used as mathematical model to represent documents as numeric vectors. We have a document D in a document vector space VS. The dimension of each document is the number of distinct features for all documents in the vector space.

We can represent document (D) in such vector space by:

$$D = \{F_{D1}, F_{D2}, \ldots, F_{Dn}\} \quad (2)$$

$$F_{D1}(VS) = \{F_1, F_2, \ldots, F_n\} \quad (3)$$

where n represents distinct features in the whole documents. Where F_{Dn} describes the weight of feature(n) in document (D). This weight represents frequency (or average frequency) as numeric value, or term frequency (TF) weight.

D. EVALUATION OF THE PROPOSED MODEL

To evaluate the proposed work (Arabic document layout analysis ''ADIR''), the previous architecture is applied for segmentation, classification, and recognition. Also, confusion matrix with detailed of the classification result will used for performance measurement.

Projection profiles, Hough transform, and nearest neighboring are used in skew estimation by calculate the skewing angle of the image [8], [9], as shown in Fig 18. Other technique estimated lines in the image and obtained skewing angle. Also, borderlines (if exist) is used to detect skewing angle using run-length algorithm [15].

Sauvola, Otsu and Niblak algorithms are the most methods used to binarize Arabic and Latin documents [16]. Therefore, local and threshold pixel intensity binarizations are used to binarize the images before the segmentation process. Segmentation extracts text paragraphs, text lines, words, or characters from scanned images. Hough transform, projection profiles, smearing, or connected components are used in the segmentation process for document layout analysis. In our work we used smearing, projection integrated with connected components to overcome the overlapped Arabic characters. Figure 19 illustrates and example of the segmented lines that used in our proposed Arabic OCR system.

Noise removing and borderline from old documents is tested in our work. The connected component is used to remove the page frame [15], [16].

Table 4 describes the analyzed results for our dataset. This analyzed data includes the documented training dataset, and documented testing dataset.

We are going to test how to analyze and recognize different categories of documents with different languages using OCR API services. Accordingly, this test used OCR engine API. Table 5 illustrates samples of analyzed documents and recognized text. We used four OCRing systems. Recognition and accuracy results of these methods are demonstrated and illustrated in table 5.

We evaluated four fine tuning OCR systems: (1) Free Abby, (2) Tesseract, (3) Novo Verus, and (4) Arabic OCR proposed in this work, illustrated in table 6.

V. ARABIC DOCUMENT RETRIEVING (ADIR)

Image datasets are collected in several domains (early printed, printed documents, journals documents, handwritten and manuscripts). Therefore, these documents their size increases, the problem of their information retrieval (IR)

TABLE 4. Analysis results of tokenization and vectorization.

Analysis Results			
Dataset	***Method Approach***	***Number of Tokens***	
		Training	***Testing***
Sientific Jounals	Segementation	204561	22203
	Recognition	8996	1000
Early Printed	Segementation	41420	5175
	Recognition	3187	355
Printed	Segementation	930364	107348
	Recognition	14803	1645
Handwritten	Segementation	48859	5894
	Recognition	3841	427

TABLE 5. Arabic documented recognition for arabic OCR.

Image Document #1	*Recognized Text #1*
	نلغات متقارب وهدفها واحد والتي من شأنها الارتفاع بالإنسان إلى العلا عن طريق الاستقامة والاعتدال الذى تحققه القيم . وهذا م يهدف إليه المعنى الاصطلاحي . ـ"للقهلصطحط إذا كان قد حدث الاتفاق على معنى القيمة لغة فإنه قد اختلف حه لها معنى ، ويرجع الاتفاق إلى أن المجتمع اتفق لغة على أن أنقيم تعنى القوة والاعتدال والاستقامة . ويرجع الاختلاف اصطلاحا إلى نظرة الإنسان إلى هذه القوة والاستقامة وهذا الاعتدال ، فهناك من عرفها بأنها: " معايير ومقاييس الفعل البشرى ومحددات السلوك ضابط أهداف النشاط الإنساني، ومسيرة البشرية في المجالات جميعا ...") ونظرا لأنها معايير ومقاييس من شأنها ضبط موازين العدالة والمساواة بين البشر. فلا يجب تركها للأهواء البشرية، حتى لا تكون محلا للعبث والتلاعب والانتقاص، فلا يمكن أن يكون الإنسان نفسه محل الفعل والتقييم والمعيارة، وأن يكون هو أيضا المعيار والمقياس في الوقت نفسه، ولا يمكن أن يتصور أن يصبح الإنسان هو الذى يجرى على فعله الخطأ والصواب أو المقياس والمعيار، والا فتح الباب للأهواء والرغبات الذاتية ومن ثم للفساد والاستغلال ، وفتح المجال لتسلط الإنسان على أخيه (١) د. عمر عبيد حسنة: مقدمة كتاب " ظاهرة العولمة رؤية نقدية "، د. بركات محمد مراد " كتاب الأمة " . قطر العدد ٨٦ عام ٢٠٠١، ص
Image Document #2	***Recognized Text #2***
	تقديم إن وظيفة الشراء قد تطورت تطوراً هائلاً خلال الخمس والعشرين سنة الماضية، ولقد كان هذا التطور نتيجة للعديد من الحقائق والوقائع التي برزت من خلال المشاهدات والتطبيق في كافة المجالات الصناعية، إذ لم يعد خافياً على أحد ما تحققه وظيفة الشراء من وفورات ضخمة يمكن أن تزيد عن أي وفورات تتحقق في أي وظيفة أخرى من وظائف المشروع نتيجة للنظم والأصول السليمة التي تقوم عليها الإدارة العلمية لشئون الشراء من جهة، وصلاحيتها للتطبيق من جهة أخرى. علاوة على ذلك ما أبرزته المنافسة الشديدة بين المنتجين والموردين في المجالات الصناعية، من الاهتمام المتزايد بتنمية الكفاءات التي تناط بها شئون المشتريات تلبية لمتطلبات النشاط الاقتصادي واحتياجاته، وإلى تطلعهم نحو حشد أعداد متزايدة من العاملين في شئون المشتريات والعارفين بسبل الارتقاء والتطور من جانب، والاتجاه بالنظر إلى وظيفة الشراء، النظرة الموضوعية حيال أية وظيفة رئيسية وهامة وقادرة بنفس الوقت على التأثير في بعض السياسات وسبل تنفيذها التي يتخذها المشروع وعلى نتائج النشاط والأعمال التي يصل إليها، مثله في ذلك مثل سائر النشاطات الأخرى في المشروع الواحد كنشاط الهندسة والإنتاج والتسويق، من جانب آخر. كذلك أن المواد والأصناف عامة والمهمات والتجهيزات من المرتكزات الهامة لمقومات الإنتاج والأنشطة التي تباشرها وتعتمد عليها المشروعات الصناعية، فالكفاية التي يحققها المشروع الصناعي تعتمد كل الاعتماد على توفير كافة الاحتياجات والإمدادات بالجودة والكمية المناسبة، وبالسعر المناسب وفي المكان والوقت المناسبين. وإن أي فشل في الوفاء بأي من هذه الالتزامات لابد وأن يضيف أعباء جديدة إلى النفقات الكلية في المشروع ومن ثم إمكانية تخفيض في الربح المتوقع، لأنه من المعروف أن أكثر من نصف ما تحققه دوائر المبيعات في الشركات الصناعية من إيرادات ينفق في توفير وتأمين الاحتياجات ومستلزمات التشغيل في العمليات الصناعية، لذلك، فإن تمكن الأفراد المسئولين من تخفيض هذا العنصر فسوف يكون بالإمكان تخفيض نفقات الإنتاج.

becomes more and more important. Consequently, documents data collection without indexing and information retrieval (IR) techniques is our aim for this framework.

An Arabic Document Retrieving (ADIR) finds segments of the printed or written text that are relevant to an information need expressed via a searching query.

The resulting IDR system will be compared using same data and experimental evaluation. The objective of this

TABLE 6. Comparative results among OCR systems.

Dataset Evaluation Results				
OCR Approach	***Scientific Journal***	***Early Printed***	***Printed***	***Hand-written***
Free Abbyy	50.00	56.10	64.50	45.00
Tesseract	66.69	75.00	78.90	55.16
Novo *	55.40	59.86	63.52	50.40
Arabic OCR	67.59	90.46	94.11	72.60

*Bilingual (Arabic and English).

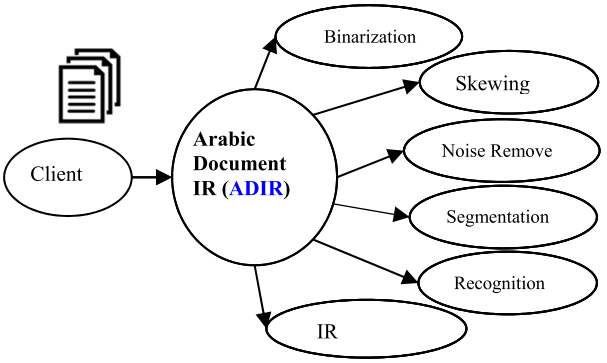


FIGURE 16. The Arabic document layout services.

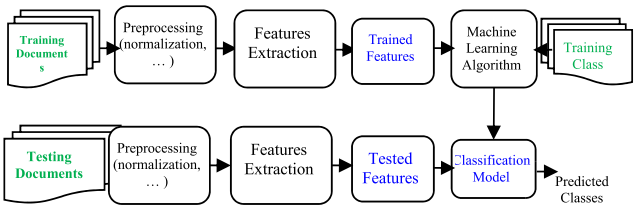
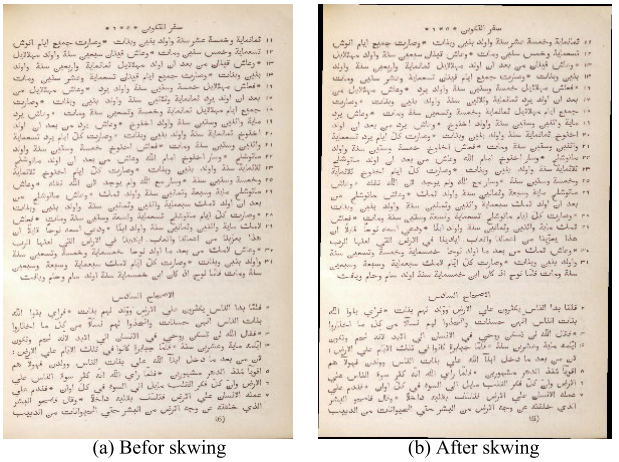


FIGURE 17. Architecture of Automated Arabic Documents Classification.



(a) Befor skwing (b) After skwing

FIGURE 18. Arabic Document (Before and After Skewing).

section is to measure the effect of the noise on the retrieval performance. To perform such objective, a state-of-the-art IDR system has been implemented and tested. At the level of

FIGURE 19. Arabic Historical Manuscripts (Before and After Segmentation).

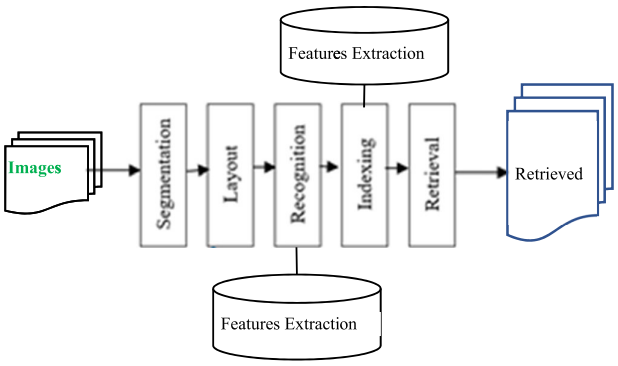


FIGURE 20. Architecture Design of the ADIR.

IR, the performance with accuracy obtained over both noisy and clean dataset using accuracy.

A. ARCHITECTURE OF IDR

The architecture design of an IDR system is composed of five modules: layout, segmentation, recognition, indexing, and retrieval, as shown in Fig. 20.

Segmentation module splits the documents stream into two segments; text segments and non-text segments as illustrated in the previous sections and in many literatures [21]–[24]. The layout module describes the document components into a stream of relevant text segments or a completely stream of non-text segments [26], [28]. As well as the ordering of each of the two streams will be illustrated. Ideally, each segment should be homogeneous in its content (relevant or non-relevant). Recognition module recognize the text-segments (relevant) into their equivalent text format (Ascii or Unicode) using OCR system [27]–[30].

Index module trying to extract semantic information for the recognized text segments. Also, the indexing approach uses the layout structure of the document into consideration with set of terms (dictionary: called bag of words). Index module gives each segment its suitable representation before retrieval task. Indexing task uses a ''finite set of terms'' or ''bag of terms'' as a dictionary (or as a database). Three models are essentially used to work with IR; the binary approach [17], the vector space approach [18], and the probabilistic approach.

The state-of-the-art systems are mostly used the vector space model. The final module (retrieval) takes a query as an input and retrieves the documents within their ranks according to their ''Retrieval Status Value (RSV)''. This module is often referred to ''on-line'' [17] retrieving.

B. THE VECTOR SPACE MODEL (VSM)

In VSM, vector represents each document, and each document's component can be associated with a term (frequency of the term). Therefore, if we have document (d_i), it can be represented by number of weighted words:

$$d_i = \{w_{1,i}, \ldots w_{T,i}\} \tag{4}$$

where: $\forall$ for all k, $w_{k,i} = f(T_k, i)$ and we can represent the common weighting function by:

$$w_{k,i} = tf_{k,I}.idf_k \tag{5}$$

in such way:

$$idf_k = \log(N/N_k) \tag{6}$$

where N_k represents number of documents in the dataset containing term k.

In case of searching a query, it can be represented in natural language, by measuring the matching between ''query vector'' and the ''documents vectors'', as a scaler product.

$$VSM(q, d_i) = \sum q_k.w_{k,I} \tag{7}$$

where $q = (q_1, \ldots, q_T)$ represents query vector.

The retrieval module takes the query, the indexed documents (resulting from all previous modules) and gives output ranking (documents are ordered according to their relevant to the query). Therefore, the retrieval task is referred as online part, while the other preceding tasks are referred as offline modules of the system. Figure 21 illustrates the structure of the retrieval module. It includes several steps: preprocessing, normalization and indexing (according to the following steps).

Space vector is computed for each document, to measure the query matching. The computed space vector should be higher for the relevant documents than the non-relevant documents.

C. HUE-SATURATION-VALUE (HSV) COLOR HISTOGRAM

The HSV color space uses a spectrum of color and illumination information. The hue space holds the color information, while the saturation and value channels determine the lightness of the color (illumination invariant features). The

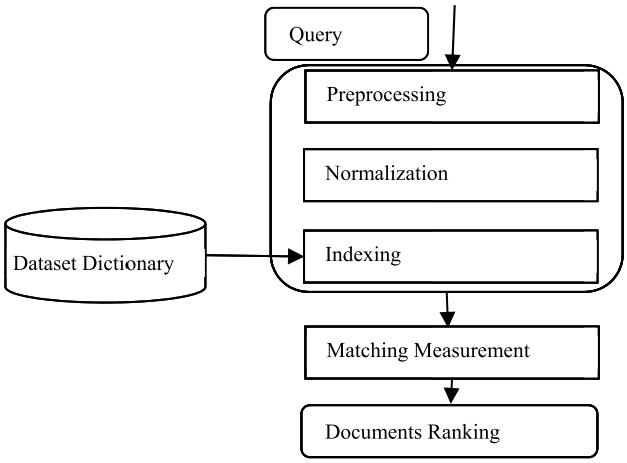


FIGURE 21. Architecture of the Retrieval Module.

hue channel can be represented as a circle with its values starting from 0 to 360. Where zero degree represents red color, 120 degree represents green color, 240 degree uses blue color, and ends with 360 degree for red color again (it starts and ends with red color).

A good features descriptor will be used in image such as pixels intensities and locations. So, textual features will be more specific representative features. Ones of the most popular of such features are used from the co-occurrence matrix. The number of co-occurrence between each pair of pixels intensities can be counted by the distance between them. So, the co-occurrence matrix (CM) finds the distance between two intensities (reference and neighbor), and the angle between them can be calculated. When the angle equal zero, this means the two pixels are on the same row (horizontal line).

The co-occurrence matrix (CM) algorithm is calculated according to the following steps:

Step 1. Convert the input manuscript image into grayscale if it is needed (gray scale or binary).
Step 2. Calculate the number of intensity levels in the manuscript image (L)
Step 3. Create a sequence matrix L x L, where both rows and columns are indexed from zero to L−1.
Step 4. Select the suitable parameters of the CM.
Step 5. Calculate the co-occurrence between each two pairs of intensity levels (pixels near to it).

To evaluate our work, information retrieval (IR) measuring is used. The variation of number of images affects to all IR measurements. Having an IR benchmark from the proposed dataset would be of great evaluation. Three main measurements are performed at three steps: segmentation, accuracy measurements, and query expansion of the Arabic OCR processes. In this evaluation, the images or the pieces of documents are represented as vectors.

The following algorithmic steps describes the information retrieving module that used for images of early printed, journals, handwritten and historical documents.

TABLE 7. Character error rate (CER) vs. word error rate (WER).

	CER	WER	Multiplier
Paragraph 1	7%	33%	4.9
Paragraph 2	20%	71%	3.6

1. Extract the historical image for Arabic documents.
   1.1 Extract the historical Arabic image using the thening algorithm.
   1.2 Classify the skeleton image into point pixels (Node set:N) and edge pixels(Edge set: E) based on information for each pixel and neighboring pixels.
   1.3 Nodes at the upper and lower sides and edges connecting sub nodes of the document image are deleted.
2. Construct the segment graph based on the historical Arabic document image.
3. Detrimine the start node and end node of the path (Vertex and Edge) in the document segmentation graph model.
4. Search and find the shortest path from the start to the end pionts to complete the segmentation of the historical Arabic document.

Table 7 illustrates the relation between some paragraphs with respect to Character Error Rate (CER) and Word Error Rate (WER).

This means $\sim$ 5% CER $\rightarrow\sim$ 25% WER. Many experts assure that 5% CER is ok for good search purposes.

D. THE STATISTICAL EXPERIMENT

We need to evaluate searching behavior through statistical testing (using query of the IR or using a search engine like Google). The IR task is to find segments of an Arabic image's dataset relevant to the searching need through a request.

What is the important statistics percentage for the suitable confidence level?

How many keywords are used for searching?

How many keywords are used for each image/document? What percent of queries are phrased as questions (quality of search)? How many words are enough to find the results of the average query? Therefore, all the retrieved images and documents are expressed in the testing with percentages.

The IR model has two inputs: a dataset of Arabic documents and a set of requests. We used RDI dataset, 93 magazines are selected one image from each magazine, and queries observations have been done on all the pages of the 93 selected articles. The first experiment of the dataset is composed from 93 articles, each article includes $\sim$10 pages (9.7 pages in average). The length of each page varies between 40-50 depending on the domain source. Their length distribution (number of lines per page) in average equals 49 lines, it includes 4575 lines, 312,511 words, 132 handwritten words, 2447 English words, 5058 Arabic poetry words, 3506 diacritic marks over words, and 27 errors for diacritics in the page sample. The statistics percentage is done with 95% with confidence level. Moreover, reference transcriptions are available. Table 8 illustrates the related statistical results of the first experiment in IR.

TABLE 8. Statistical results for IR of the OCR system.

Item Description	*Total #*	*Average*	*Percentage*
No of pages	903	9.7	
No of lines	4575	49.2	
No of words	312511	3360	
Handwritten	132	1.1	0.04%
English words	2447	26.3	0.8%
Arabic poetry words	5058	54.4	1.6%
Diacritics marks over words	3506	37.7	1.1%
Errors of diacritics in pages	27	0.12	12.4%

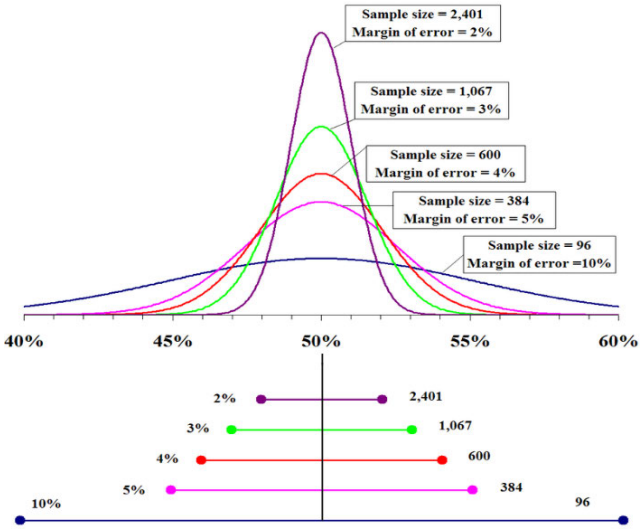


FIGURE 22. Experimental statistics percentage with 95% with confidence level.

Figure 22 illustrates the t-distribution w.r.t normal distribution. It shows that the absolute errors are quite trivial, however the relative errors raise higher at the ends.

For each text segment, two different transcriptions are available: the first one is being produced by the expert annotators manually and the second is obtained with OCR systems. The first transcription is considered as clean text, even if it is affected by WER of ≅ 5%. The OCR output has a 40% WER and plays the role of a noisy text. There are approximately 93 articles in the manual segmentation produced in the dataset.

Exp 2: Keywords selection to answer the second question

To test the quality of search, we selected one essay from each magazine (93), then we distributed the 93 essays on few linguistics. Each one selects about 5 key words (KWs) from each essay to be considered as search items. Our findings are as shown in table 9 as a keywords IR Experiments

Exp 3: Quality of ADIR for selected KWs

In case of testing the quality of search, the performance of the ADIR process used Arabic keywords for searching is shown in table 9. The effect of errors on the IR task can be measured relative to the number of keywords. As shown in the table, the performance is improved relative to the number of keywords. We searched – using these KWs- through the given essays (16 essays have been finished so far) 1 essay from a different domain, our findings are as illustrated in table 10.

TABLE 9. Statistical results for IR of the Lingistic experts.

Item Description	*Total #*
No of pages (Articles)	**93**
No of keywords in the 93 articles without repitation	**451**
No of keywords in the 93 articles with repitation	**4500**
No of pages (articles) that include one keyword or more	

TABLE 10. 3rd Experiment: ADIR results for IR relative to KYs.

Item Description	*Original Documents*	*Errors*	*Percentage*
Total no of the keywords errors	624	22	3.5%
No of documents that not identified	206	6	2.91%
No of documents that not identified using one KW	16	1	6.25%
No of documents that not identified using 3 KWs	16	0	0.00%
No of documents that not identified using 5 KWs	16	0	0.00%

FIGURE 23. Result of the IR using keyword "محمد" that is extracted from the obtained dataset.

Figure 23 has shown the effect of the IR for searching using keyword "محمد" that is extracted from the obtained dataset. However, the selection of the noise may contain error. So, IR is very immune for errors

From Google: Aver. No. of words/query https://moz.com/blog/state-of-searcher-behavior-revealed

- a typical searcher uses **about** 3 **words** in their queries.
- Desktop users: can go to queries of 6 words or more, see figure 24.
- The answer is shockingly big: a full 18% of searches lead to a change in the search query!

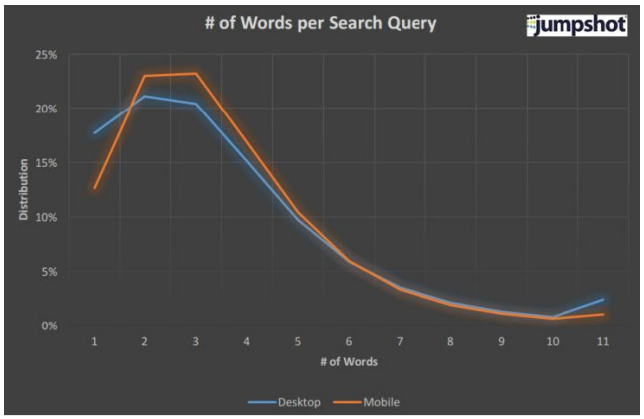


FIGURE 24. Distribution relative to the number of retrieved Arabic words.

To conclude Experiment 3

- ➢ WER < 25-30% or CER < 5% is accepted for OCR
- ➢ Even if less accuracy is collected for noisy documents, still many other issues help:
  1. When having multiple existence in the same page, the chance to hit this page will still exist
  2. Stemming before indexing
  3. The same KW comes in many pages for the same article, so if one page is lost, other page(s) will hit
  4. On the average the query has 3 words +
  5. The user even using Google may reconsider other KWs

At the ADIR stage, the performance obtained better results (∼75% improvement in accuracy) with both the standard dataset and other documents dataset.

The results obtained in table 11 are comparable with those published at Arabic OCR systems (where values vary from 50.25% and 78.50% for same documents data). Such results are based on 93 articles of pages.

Handwriting, decorative lines, or diacritics

E. THE OVERALL MANUSCRIPT RETRIEVING DOCUMENTS

Many of researches use text-based searching models, such as Bing and Google by using few of keywords related to the entire contents we need to retrieve. This section introduces to new strategy of searching scenario using images of the manuscripts instead of the text as a query. Therefore, the proposed searching strategy will use meta descriptor searching approach. So, the proposed work finds and returns near-identical matches of the query manuscript images. Accordingly, some sorted algorithms can be used to extract features or discriptors that represent manuscript images from the dataset.

Tagging and classification methodes are used in manuscript and image searching to find the matched or similar stored manuscripts. Such methodes need more time in addition to manual efforts. Therefore, another method is needed to enhance the searching speed and to visualize descriptors. One of the most searching methods for manuscript retrieving is based on text searching using few

TABLE 11. Comparative test among arabic OCRs.

Item Description	*Items*	*Experiments*		
		#1	#2	#3
Features	Arabic	√	√	√
	Handwriting	√	√	√
	Non-normalized	√	√	
	Latins	√		
	Symbols	√		
	Numbers	√		
OCR System	Proposed OCR	83%	90%	94%
	FRE (Abbyy)	50%	56%	59%
	Tesseract	66%	76%	78%
	Novo	51.5%	59%	63.5%

FIGURE 25. Seven errors in printed text due to 90- Arabic words with diacritics. At least 4 handwriting or decorative faults.

keywords related to the objects contents of the manuscript. The second searching method is related to manuscript itself as an image, by quantify the contents to make it searchable. Therefore, three strategies can be used for manuscript searching: (1) Manuscript or meta descriptor strategy, (2) Query strategy, and (3) Fusion/Hybrid Strategy. In the manuscript strategy, the contents of the images are used to perform searching method. Accordingly, the manuscript is analyzed, quatified, indexed and stored with similar images. So, the searching method is relying strictly on the features contents of the manuscript without textual annotations. In the second strateg, tags description is needed to prsent each manuscript to the local/global host based on the query. The fusion/hybrid strategy is based on the two approaches; contextual keywords along with a serach by image itself.

A scalable Arabic manuscript hashing search engine using computer vision and hashing algorithms are used to: (1) Analyize and quantify the manuscript contents into an uniquely single integer, (2) Find similarity or duplicate in a dataset of manuscripts using their computed hashes indexes. Additional method is used to improve the searching performance based on VP-Tree (Vantage Point Trees) [25]. Accordingly, a text search engine is implemnted and used to return by similar manuscripts results based on the content of the entire query manuscript.

The fusion/hybrid strategy is based on employing the deep learning using hand-crafted features, and is implemented and tested. The limited annotated dataset has facilitated progress of the analysis, features extraction and the testing.

At the testing phase, we used three dataset sources. The first source is from university of Pennsylvania's rare book [27]. We collected 100 manuscript images for annotation (it needed). Our vision is going to maximize the diversity of the Arabic manuscript dataset images types and domains, with document degradation, script language, number of lines, number of writers, non-textual elements, etc. Some of these manuscripts include multiple of information stacked horizontally or vertically or with oriented angle.

The second source for our dataset from printed, early printed Arabic documents. It includes at least 100 of images. However, the third source includes calligraphy Arabic documents with different writers.

VI. CONCLUSION

This paper introduced details of Arabic document layout analysis about ADIR services. Therefore, detailed of the preprocessing to collect dataset with filtering, removing unrelated documents and annotating the documents are discussed. Accordingly, we have introduced the ADIR services to provide and support an infrastructure to support creativity of datasets and OCRing systems. This infrastructure uses HTTP protocol to transport messages between users and services over the internet and is based on the use of URIs to refer to dataset resources.

We have presented in this paper, a proposed model to analyze and classify Arabic documents. The real application domain includes many of dataset in different domains. The proposed solution tested with four types of corpora for Arabic documents. The proposed model has tested with different three different OCRing services, the output results indicates that our model increased the accuracy of the OCRing segmentation and recognition tasks in all the used datasets based on the document model.

Documents data collection was our aim for this framework. An image document retrieving with ADIR finds segments of the text that are relevant to an information need expressed via a searching query. The results of the ADIR system are compared using same data and experimental evaluation. The objective of this ADIR is to measure the effect of the noise on the retrieval performance. At the level of IR, the performance with accuracy obtained over both noisy and clean dataset using accuracy standard relative of the three strategies. The CNN was used for Arabic handwritten letters classification. The proposed CNN trained and tested using the whole 168,000 Arabic handwritten letters. The dataset is segmented in training set with 13,440 letters for 48 images per class label, and testing set with 3,360 letters to 120 images per class label. In an experimental test, we showed that our results outperform 100% classification accuracy rate on testing images. When in doubt, the general rule is more data we had, is the better. Moreover, depending on the corpus size (documents content), training can take several hours or even days, but fortunately, we can store the analyzed data and extracted features on a storage drive. This way we do not have to do the analyzed tasks of model training every time we need to use it.

We are combined two approaches, minimize OCR errors and acquire text inquiry for IR. The two approaches are tested, evaluated and judged tested in three different experimental domains. Some difficulties such as understanding the concept and the meaning of scanned images and related definitions need to additional elaboration. The idea behind that is the archiving meanings should appear in similar context. Therefore, official repository for semantic information for domains' contents and terminologies definitions should be involved using ontological additional works. Although, multilingual documents will be included in the IR future work.

ACKNOWLEDGMENT

This work was supported by the National Plan for Science, Technology and Innovation (MAARIFAH) – King Abdulaziz City for Science and Technology -the Kingdom of Saudi Arabia under Award 11-INF-1997-03. The authors thank the Science and Technology Unit, King Abdulaziz University for technical support.

REFERENCES

[1] D. Lopresti and B. Lamiroy, "Document analysis research in the year 2021," in *Proc. Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst.*, in Lecture Notes in Computer Science, vol. 6703, 2011, pp. 264–274.

[2] D. Doermann and K. Tombre, Eds., *Handbook of Document Image Processing and Recognition*. London, U.K.: Springer, 2014.

[3] M. Wursch, R. Ingold, and M. Liwicki, "SDK reinvented: Document image analysis methods as RESTful Web services," in *Proc. 12th IAPR Workshop Document Anal. Syst. (DAS)*, Apr. 2016, pp. 90–95.

[4] B. Lamiroy, "DAE-NG: A shareable and open document image annotation data framework," in *Proc. 1st Int. Workshop Open Services Tools Document Anal., 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, 2017, pp. 31–34.

[5] F. Stahlberg and S. Vogel, "QATIP—An optical character recognition system for arabic heritage collections in libraries," in *Proc. 12th IAPR Workshop Document Anal. Syst. (DAS)*, Apr. 2016, pp. 168–173.

[6] K. Santosh, *Document Image Analysis: Current Trens and Challenges in Graphics Recognition*. New York, NY, USA: Springer, 2018.

[7] H. Al-Barhamtoshy, M. Khemakhem, K. Jambi, F. Essa, A. Fattouh, and A. Al-Ghamdi, "Universal metadata repository for document analysis and recognition," in *Proc. IEEE/ACS 13th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2016, pp. 1–6.

[8] H. M. Al-Barhamtoshy, "Towards large scale image similarity discovery model," in *Proc. 2nd Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, Monastir, Tunisia, Mar. 2016, pp. 1–9.

[9] A. M. Hesham, S. Abdou, A. Badr, M. Rashwan, and H. M. Al-Barhamtoshy, "A zone classification approach for arabic documents using hybrid features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 7, pp. 158–162, 2016.

[10] K. C. Santosh, "G-DICE: Graph mining-based document information content exploitation," *Int. J. Document Anal. Recognit.*, vol. 18, no. 4, pp. 337–355, Dec. 2015.

[11] A. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," in *Proc. 1st Int. Workshop Arabic Script Anal. Recognit. (ASAR)*, 2017, pp. 114–118.

[12] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis, and opinion mining," in *Proc. LREc*, 2010, pp. 1320–1326.

[13] H. Al-Barhamtoshy and A. Al-Ghamdi, "An OCR Web services system for arabic calligraphy documents," *Int. J. Eng. Technol.*, vol. 8, no. 1.11, pp. 16–24, Mar. 2019.

[14] H. Al-Barhamtoshy, K. Thabit, and B. Bal-Aziz, "Arabic morphology template grammar—Based," in *Proc. 7th Conf. Lang. Eng.* Cairo, Egypt: Ain Shams Univ., Dec. 2007, pp. 216–234.

[15] H. M. Al-Barhamtoshy *et al.*, "Arabic calligraphy, typewritten and handwritten using OCR system," *Biotech. Res. Commun.*, vol. 12, no. 2, pp. 283–296, Apr./Jun. 2019, doi: 10.21786/bbrc/12.2/11.
[16] S. S. A. Mohamed, M. A. A. Rashwan, S. M. Abdou, and H. M. Al-Barhamtoshy, "Patch-based document denoising," in *Proc. Int. Japan-Africa Conf. Electron., Commun. Comput. (JAC-ECC)*, Dec. 2018, pp. 160–164.
[17] N. Naji, J. Savoy, and L. Dolamic. (2011). *Information Retrieval With a Noisy Text Corpus (OCR)*. [Online]. Available: https://www.researchgate.net/publication/287379560_Information_retrieval_with_a_noisy_text_corpus_OCR
[18] B. M. Schmidt, "Plot arceology: A vector-space model of narrative structure," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 1667–1672.
[19] L. Ma, C. Long, L. Duan, X. Zhang, Y. Li, and Q. Zhao, "Segmentation and recognition for historical tibetan document images," *IEEE Access*, vol. 8, pp. 52641–52651, 2020.
[20] A. Dulla, "A dataset of warped historical arabic documents," in *Proc. 9th Int. Conf. Pattern Recognit. Syst. (ICPRS)*, Valparaiso, Chile, May 2018, pp. 22–24.
[21] H. Al-Barhamtoshy, K. Jambi, H. Ahmed, S. Mohamed, M. Rashwan, and S. Abdou, "An OCR system for arabic calligraphy documents," *Int. J. Eng. Technol.*, vol. 8, no. 1.11, pp. 9–16, Mar. 2019. [Online]. Available: http://www.sciencepubco.com/index.php/IJET, doi: 10.14419/ijet.v8i1.11.28083.
[22] H. M. Al-Barhamtoshy and A. S. Al-Ghamdi, "A comprehensive framework for ocr Web services system for arabic calligraphy documents," *Int. J. Eng. Technol.*, vol. 8, no. 1.11, pp. 16–24, Mar. 2019. [Online]. Available: http://www.sciencepubco.com/index.php/IJET, doi: 10.14419/ijet.v8i1.11.28084.
[23] K. Jambi, H. Al-Barhamtoshy, A. Fattouh, A. Al-Ghamdi, F. Eassa, and M. Khemakhem, "An open architecture for enhancing performance of complex OCR applications," *Int. J. Eng. Technol.*, vol. 8, no. 1.11, pp. 154–157, 2019. [Online]. Available: http://www.sciencepubco.com/index.php/IJET, doi: 10.14419/ijet.v8i1.11.28188.
[24] H. M. Al-Barhamtoshy, K. M. Jambi, H. Ahmed, S. Mohamed, S. M. Abdo, and M. A. Rashwan, "Arabic calligraphy, typewritten and handwritten using OCR system," *Biosci. Biotechnol. Res. Commun.* vol. 12, no. 2, Apr./Jun. 2019. [Online]. Available: http://www.bbrc.in, doi: 10.21786/bbrc/12.2/11.
[25] J. Vankerschaver, R. Kern, S. J. Kern, P. Zahemszky, C. Mueller, and R. Cardwell, "Searching efficiently through genomic," presented at the 69th Annu. Meeting Amer. Soc. Hum. Genet., Houston, TX, USA, Oct. 2019. [Online]. Available: http://www.enthought.com/wp-content/uploads/2019/08/ASHG_Poster_PgmNr_1640_-10-10-19-2.pdf
[26] C. K. Savitha and P. J. Antony, "Machine learning approaches for recognition of offline Tulu handwritten scripts," *J. Phys., Conf. Ser.*, vol. 1142, Nov. 2018, Art. no. 012005.
[27] A. Abeysinghe and A. Abeysinghe, "Use of neural networks in archaeology: Preservation of assamese manuscripts," in *Proc. Int. Seminar Assamese Culture Heritage*, 2018, p. 27.
[28] *Penn in Hand: Selected Manuscripts*. Accessed: Feb. 1, 2020. [Online]. Available: http://dla.library.upenn.edu/dla/medren/search.html?fq=collection facet:"IndicManuscripts"

HASSANIN M. AL-BARHAMTOSHY (Fellow, IEEE) received the B.S. degree in electronic and communication engineering from Cairo University, in 1978, and the M.S. and Ph.D. degree in systems and computers engineering from Al-Azhar University, Cairo, in 1985 and 1992, respectively. From 1992 to 1997, he was an Assistant Professor with the Department of Systems and Computer Engineering, Al-Azhar University. From 1996 to 1997, he was an Assistant Professor of computer science with KAU University, Jeddah, Saudi Arabia. From 1998 to 2002, he was an Associate Professor. Since 2003, he has been a Professor with the Department of Computer Science and Information Technology, Faculty of Computing and Information Technology, KAU University. He is currently a Professor with the Faculty of Computing and Information Technology, IT Department, KAU University. His research interests include language processing and machine translation, image processing, Arabic optical character recognition, intelligent systems, and speech processing. He is a member of the review committee in several conferences and journals in the language engineering fields. He is the Principal Investigator and a Co-Principal Investigator of several research projects in the areas of language engineering.

KAMAL M. JAMBI received the M.S. degree in computer science from Michigan State University, East Lansing, in 1986, and the Ph.D. degree in computer science from the Illinois Institute of Technology, Chicago, in 1991. He was a PI in many research projects from KACST and KAU. He was a Former Vice Dean of the Graduate Studies & Scientific Research with FCIT. He is currently a Professor of computer science with King Abdulaziz University, Jeddah, Saudi Arabia. He is a Professor with the Computer Science Department. His research interests include AI, deep learning, blockchain, resilience, and bigdata. He is a member of the Saudi Computer Society and the Pattern Recognition Society. He is a member of the review committee in several conferences and journals in the artificial intelligence fields. He is the Principal Investigator of several research projects in the areas of computer sciences and technologies.

SHERIF M. ABDOU received the B.Sc. and M.Sc. degrees in computer science and automatic control from the University of Alexandria, Egypt, in 1993 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Miami, USA, in 2003. He joined BBN Technologies as a Senior Staff Scientist, in 2003, in the Arabic language team of the Ears project to provide affordable reusable speech-to-text decoding for the Defence Advanced Research Projects Agency, DARPA. He was appointed as the Research and Development Manager of the Research and Development International (RDI) Company, in 2005, where he is leading a team to develop several products for natural language processing, computer aided language learning, speech recognition, speech syntheses, optical character recognition, handwriting recognition with special focus on the technologies of the Arabic language. He joined the Information Technology Department, Faculty of Computers and Information, Cairo University, as an Assistant Professor, in 2005. He is one of the holders of the patent Systems and Methods for Quran Recitations Rules: HAFSS. He is a member of the review committee in several conferences and journals in the HLT fields. He is the Principal Investigator and Co-Principal Investigator of several research projects in the areas of language learning, virtual tutors, web monitoring and intelligent contact canters.

MOHSEN A. RASHWAN is currently a Professor of communications with the Department of Electronics and Communications, Faculty of Engineering, Cairo University, Egypt. He had over 100 papers are published in international proceedings and conferences. He had over 70 theses under his supervision (finished and current): 26 Ph.D.'s and 47 M.Sc.'s. Many of his ex-postgraduate (M.Sc. and Ph.D.) students are currently recruited in the core of top world hi-tech companies and research centers, such as IBM-WRC, Lucent Technologies, Microsoft, and so on. Over 350 graduation projects are realized under his supervision with the Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, in different applications, such as the digital processing of speech, audio, image, and video, as well as pattern recognition and/or classification, OCR, document analysis, biometry, software, and hardware design. Many prizes have been awarded to some of those projects. He has a grant of research and development of many projects from the Information Technology Academia Collaboration (ITAC) program initiated by the Information Technology Industry Development Agency (ITIDA); www.ITIDA.gov.eg for RDI exclusively to produce its Arabic off-line OCR. From April 2010 to June 2011, he was the Principal Investigator of this project. He had Mediterranean Arabic Language and Speech Technologies (MEDAR). This project runs under the EU's FP7 Research and Development grants program. This project is shared by 15 partners from 11 ME and European countries, from February 2008 to August 2010.