International Journal of Engineering & Technology, 8 (1.11) (2019) 16-24

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Comprehensive Framework for OCR Web Services System for Arabic Calligraphy Documents

Hassanin M. Al-Barhamtoshy, Abdullah S. Al-Ghamdi

Computing and Information Technology, King Abdulaziz University, KAU, Jeddah, Saudi Arabia
**E-mail: hassanin, abdmalaise @kau.edu.sa}*

Abstract

This paper describes document layout analysis web services approach for OCR systems, in case of integrate with web-based applications using SOAP and REST interfaces. The proposed solution provides accessing way to use different OCR systems. Therefore, these web services are implemented using SOAP and REST interfaces through HTTP or HTTPS requests. Consequently, different developers can communicate with each other's without time consuming to customize code implementation, operating system barriers, and programming language conditions.

The scientific scope of this paper focuses on three objectives:
(1) The document categories on which they are included in the dataset, (2) The related algorithms that are used in the level of document analysis, and (3) The Arabic document image segmentation algorithms they are used. Consequently, the connected components method is used to remove page frame in the old and calligraphy documents. Also, shadow noises in the old and historical documents are removed using the adapted sparse algorithm.
This paper discusses a number of the major areas where OCR web services have been working comprehensively: in supporting document analysis and OCR service-oriented architecture computing. Using the OCR web services approaches, we are dealing with heterogeneous large scale documents with wide varying structured category. Furthermore, there could be multipage document with different languages. Accordingly, the language domain will be identified within the language script specification module.

Keywords: *Arabic; Document analysis, connected component; sparse; segmentation; OCR web services.*

1. Introduction and related works

The OCR web services is state of the art in the field of document analysis. Therefore, the OCR web services for document accessing and remote method invocation are still uncharted in the document layout analysis.

The document outline analysis is the method of classifying and labeling the zones of the image documents into a segments of text documents. Any OCR system involves the segmentation of text regions from non-textual ones. But, document's text regions play diverse its logical parts inside the manuscript (title, subtitle, notes, cross-references, etc.) and this kind of semantic labeling is the opportunity of the layout analysis.

Recent submissions that used Arabic text segmentation and Arabic OCR systems are discussed in [1]. Other published applications to recognize Arabic handwritten is illustrated in

[2]. A wide range of document analysis, layouts processing to locate text blocks and none text blocks, and separated between

them have been presented in [3]. Script identification has been a real challenge in OCR and information retrieval systems [4]. The most state of the art papers are published into the OCR machine printed character domain. Moreover, segmentation process leads to errors more than the other processes in document analysis and processing [2].

Text classification addresses the problem of document analysis into "classifying modern machine printed text, handwritten text and historical typewritten text from degraded noisy document" [5]. Therefore, text classification approach based on vector is proposed in [5]. The text line is classified using SVM in vector space.
An OCR for multilingual documents (Amazigh-French) has been proposed in [6]. Hence, Amazigh writing transcription methods are employed using Latin or Arabic alphabet. Accordingly, such as paper [6] focused on Amazigh population by transcribed in Latin.

Another Arabic text recognition system is presented using statistical features and mono HMMs by extracting set of simple features from one pixel with sliding window [7]. Then, it injects such feature vectors to the HMM. Holistic Arabic printed word recognizer is intro-

duced along with discrete Markov classifier, HMM toolkit (HTK), and discrete cosine transform (DCT) [8]. Five fonts are used, having size of 14 points with plain style. Additional details, technical points and paper analysis are presented there [8].

Wurch et al. [9] present RESTFull web services with SDK for document image analysis. They introduce DIVAServices framework to access analytical methods. Their framework allows a developer to use algorithms with related data parameters that result integration for further processing.

Cholia et al. [10] provide an accessing methodology through command line tool to their scientific computing center via RESTFull web services in high performance applications.
Consequently, the OCR web services are still in early stage for document layout analysis. Lamiroy and Loperesti introduce a set of algorithms and document dataset with a SOAP web service into their document analysis and exploitation (DAE) system [11].

The paper is organized as the following description. Section
II introduces an overview of dataset description, and categories of documents, and multi classification of OCR system and related definitions. Section III gives the proposed solution of the Arabic OCR system with preprocessing details description (binarization, skewing, frame removing, and segmentation modules). Section IV provides language model and lexicon
building with Arabic OCR dataset description. Section V introduces the features extraction and HMM decoding description. Performance evaluation results will be discussed in section VI. Section VII summarizes conclusion and feature works.

2. Dataset description and document category

One of the challenges in the domain of image layout analysis and recognition systems is the "annotated datasets creation". This dataset can share parameters with different methods in the layout analysis and the recognition systems. The proposed dataset is used to assess the performance of different OCR systems.

The documents to be studied within the proposed datasets are completed of different fields under different formats styles (document styles, sizes, colors …). Fig. 1 shows samples of the Arabic documents. Accordingly, the variants in such documents in syntax, styles, and colors during the analytic processes are needed of the proposed system. The proposed dataset categorizes the major types of documents from the oldest or historical documents to the earlier (modern documents). In fact, many categories have been illustrated on both ancient and modern documents.

Fig. 1: Samples of Document Categories

Arabic historical documents are classified as one of the most important documents that includes historical, political, and ancient information over the world archives. Figure 2 shows two examples of Arabic historical documents and multi script (Arabic/English) with different kinds of orientation.

Fig. 2: Calligraphy and Handwritten Ancient Arabic Documents

3. Document layout analysis and multi-language classification

Analyzing the layout of documents is difficult, particularly in calligraphy, typewritten, handwritten, historical images, and manuscript notations. Any documents can be classified based on its category; either printed, typewritten, handwritten, or scripts and manuscripts

documents. Such documents require dedicated preprocessing modules because of some common properties, structure, clearing irrelevant objects or noises, language direction…etc. Therefore, the images require particular pre-processing procedures because of some familiar properties such identifying unrelated objects or noises, language detection and language direction …etc. The proposed framework of the OCR system takes into consideration the different categories of such documents categories. Fig. 3 illustrates a flow diagram of multi-classification that can be modified in the proposed OCR technology.

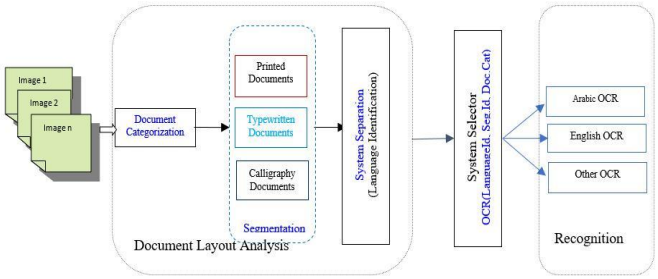

Fig. 3: OCR System Framework with Multi Document Classification

Generally developing an OCR system framework for multi-script language documents is more difficult than a single-script OCR, due to features necessary on the language models, structures, properties, styles, and nature of writing. The first module in the layout analysis is to organize between text and non-text regions. Therefore, given a document image, a decomposition of such document into smaller re-gions. Thus, these regions are classified as including text or non-text elements. Such text regions are fed to the second module of the OCR system, in order to declare category of the document (printed, calligraphy … or typewritten). The proposed solution is dealing with heterogeneous large scale structured and non-structured categories. Furthermore, multipage documents with different languages could be accessed. Accordingly, the language domain will be identified within the language script specification module.

3.1. Documents categories

Two types of layout analysis to define document contents description are required; physical and logical layouts. The physical layout de-scribes the nature of the contents of the document such as text and non-text regions. The logical layout illustrates the function of the page contents (header, main body, title, subtitle, …, footnote) [12]. Table 1 shows the detailed characteristics of dataset to work with.

Table 1: Detailed dataset description

Image Type	Total Images	Training(Developing)	Testing(Evaluation)	Dataset Name	Language	Data Range	Availability
Handwritten	1500	-	-	CDAR	English	Early images	Proprietary
Handwritten	1539	-	-	IAM	English	Early images	Public
Handwritten	7447	-	-	LDC	Arabic	Early	Private
Early Print	20	13	7	RDI	Arabic	Early images	Public
Calligraphy	10	7	3	Arabic OCR	Arabic	Before 1000	Public
Books	20	13	7	RDI	Arabic	After 1980	Public
Handwritten	80			ISI	Bengali		Upon Request
Handwritten	540			AmirKabir	Farsi		Unknown
Handwritten	2200			IFN/ENIT	Arabic		Public
Printed				ALTC	Arabic		Upon Request

3.2. Language Detection and Identification

The goal is to improve the OCR accuracy by involving auto selection of the OCR services. To enable this goal, the layout module aims to:

1. Identifying the segment image category.
2. Detecting and segmenting text and non-text regions.
3. Expect the language of each segmented region to invoke related language model.

A script is a visual representation or an organized arrangement of distinct graphic characters in precise language patterns known as the alphabet of a language [12]. Grapheme, allograph and glyph are needed to study the script of a language identification, and they contrib-ute to the script formation, as well as designing fonts [12]. Any OCR can be carried out using one of the following steps:

1. Building a general OCR that recognizes all words and characters of the alphabets in all possible languages.
2. Building language separation module (language model) to identify each single script with related OCR engine.

3.3. Layout Analysis and Segmentation Algorithms

The segmentation algorithm can be classified into three categories (techniques) [13]. They are used to work with segmentation survey: Top-down approach, bottom-up approach, and hybrid approach. The top-down approach starts from the entire document and attempt to partition it into segments or sub-segments. The bottom-up approach starts from a small elements and attempt to agglomerate such ele-ments into bigger one to the scale of the entire page. Accordingly, three main scales are used: pixels scale, connected components scale, and user defined scale (patches).

Before we are going through the OCR classification and recognition tasks, it is reasonable to select a suitable algorithms for segmenting such entire document/segment. The overall classification can come from the way of the algorithm itself, such as how segment a specific regions with projection profiles (horizontal and vertical projection). Also, it come from the parameters passed to the algorithm, depending on the document style (limitation). Therefore, both techniques and limitation are combined together to produce the third level [3].

The algorithm in the first type, we aiming to segment a specific region without any training. There are three subtypes:

(1) Algorithms used projection profiles. (2) Algorithms used filter and noise removing techniques. (3) Algorithms identified straight line or square borders.

3.4. Arabic document layout analysis

This paper aims to develop suitable OCR system by selecting a tuned version of the pre-processing modules (especially for Historical and Calligraphy documents). The Arabic heritage archiving organizations has a storage of millions of historical Arabic documents. Having an Arabic OCR system that can work with this category of documents is

one of the high priorities. Calligraphy and historical documents strongly differ of the other early documents, as it contains a lot of challenges i.e., there is a low layout formatting requirements, so their physical structure is thus harder to extract. In addition, calligraphy and historical documents are of low quality, due to aging or faint typing. They include various disturbing elements such as holes, spots, writing fragmentation artifacts. Pages may include narrow spaced lines with overlapping and touching components. In case of Arabic historical documents dots and diacritics may cause a lot of problems, because some dots are very far away from their components. Fig. 4 shows the sequence of the pre-processing operations used in any OCR system for documents analysis.

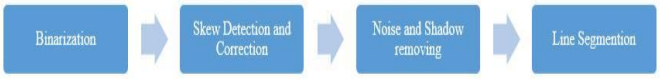

Fig. 4: Preprocessing of the OCR System

Some Arabic calligraphy and Arabic historical documents are different from those in other categories. The difference is that the existing of rectangular frames around the texts. These frames need to be extracted or removed. Consequently, in order to remove such frames, we first extract the horizontal and vertical lines. Figure 5 shows the block diagram of the noise removal module. Figure 6 (a & b) show sample processed image after noise removal.

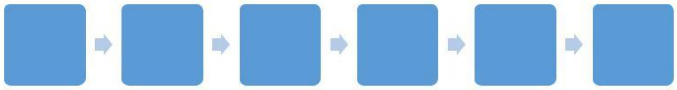

Fig. 5: Frame detection and removing

In addition, sparse algorithm is used to remove shadow noise that is common effect in old and historical documents

[8]. The main advantages of this algorithm are its ability to use a learned prior knowledge in noise removal process.

(a) **Before frame removal** (b) **After frame removal**

Fig. 6: Frame detection and removing for Arabic document

4. Web services infrastructure and related tasks

Therefore, the document analysis and OCR services are designed to be used in distributed systems, they will be established using collections of lightly coupled services that can be discovered and connected with each other to provide enhanced services. The document analysis and OCR services are realized through the use of web services technology. The proposed solution offers a common vision of services creating them universally and cooperate to succeed the levels of documents heterogeneity composition.
Also, to deal with the problem of diverse organizations services with different middleware internally it is possible for one organization to use document analysis and another to use OCR subservices, then both services expose interfaces using the global interoperability of the web.
Figure 7 encapsulates the main components almost the communication architecture in which OCR web services operate: any OCR web service is acknowledged by a uniform resource identifier (URI) and can be accessed by consumers using messages scripted in XML. SOAP is used to summarize these messages and transmit them over HTTP or another protocol (e.g. TCP or SMTP). OCR web service deploys service descriptions to specify the interface and other aspects of the service for the benefit of potential consumers. Such OCR web services can implemented using SOAP interface and can be retrieved through API by HTTP or HTTPS requirements. The top layer of the figure illustrates the following:

1. Layout analysis of the OCR Web services may be built on top of other web services.
2. Some particular web services provide general tasks required for the process of a great amount of other OCR web services. They include, document category, language identification and segmentation, all of which are discussed. An interface is needed to communicate with OCR web

services. Therefore, the OCR web service description includes two parts, communication method part (SOAP over HTTP), and the URI of the OCR web service itself. So, such description is written in XML format. The OCR-WSDL is abbreviation of OCR web services description language, it is used for the OCR web service description with an XML schema. It contains the service name, service type, message, interface, bindings and services.

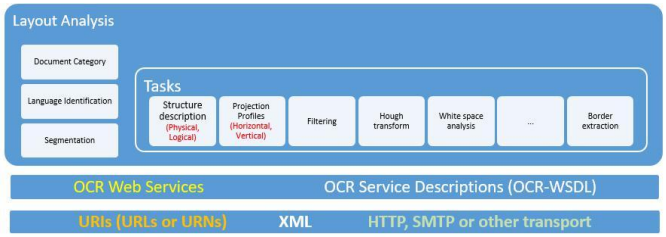

Fig. 7: Layout Analysis Web Services Infrastructure and related Tasks

An interface is needed to communicate with OCR web services. Therefore, the OCR web service description includes two parts, communication method part (SOAP over HTTP), and the URI of the OCR web service itself. So, such description is written in XML format. The OCR-WSDL is abbreviation of OCR web services description language, it is used for the OCR web service description with an XML schema. It contains the service name, service type, message, interface, bindings and services. Apply a sliding window of 11 pixels in width along the writing direction.

4.1. Proposed Architecture Scenario

The proposed framework facilitates a mashup methodology
[15] by combining two or more services available in the distributed environment. The mashup methodology relies on the available services with well-defined interfaces coupled with an open innovation community and make them available to others for further development.
Then, the "Grid technology" is used to refer to enable the sharing of the document analysis and OCR resources such as documents' datasets, images files, computers, software library, data and devices (such as scanner and camera) on a very large scale. The document analysis and the OCR resources are shared by sets of developers in different organizations requiring large numbers of computers to solve them, either by the sharing of datasets and services or by the sharing of computing power. These resources are necessarily supported by heterogeneous computer hardware, operating systems, API's, programming languages and applications. As an example to develop features the document analysis and OCR web services, the following steps are needed:

- Datasets are collected and annotated by experts and scientific instruments;
- The dataset are stored in archives at separate sites whose locations can be in different places throughout the world;
- The dataset is managed by teams of experts belonging to discrete societies;
- Huge of data with quantity (terabytes or petabytes) of raw data is generated from the experts and scientific instruments;
- Computing tools are used to investigate and make reviews of the raw dataset, to classify, calibrate and categorize the raw datasets into categories base on its language and field of domain.

The fact that any dataset is processed at many different sites offers an integral parallelism that effectively divides the huge task being undertaken. From the above characteristics, the following requests are derived:

R1: Remote access to resources – that is, to the required documents information in the datasets.
R2: Processing of datasets at the site where it are warehoused and accomplished, either when they are collected or in response to a request.

R3: The OCR resource manager of a dataset collection should be able to create service instances to deal with the particular section of dataset required, just as in the distributed object model, where servants are created whenever they are needed to handle different resources managed by a service.
R4: Metadata to describe:

- Characteristics of the dataset in an collection- i.e., for historical or calligraphy documents, the area field of the document, the date and time collected and the instruments used for scanning;
- Characteristics of the OCR web service managing that data- i.e., its cost, its geographic location, its publisher or its load or space available.

4.2. The OCR SOAP and REST Paradigms

The document analysis APIs, and the OCR web services' APIs can be used as a cloud based services that provide an interfaces (SOAP and/or REST) to integrate layout analysis and OCR technologies into a new software products, mobile devices or another web services. Therefore, the OCR SOAP and REST parameters description are illustrated in Figures 8 and 9. The SOAP and REST approaches of the OCR web services can support many recognized languages with different formats (PDF, TIF, DOC, BMP, PCX, JPEG, PNG, and GIF as input). The output of the recognized text can be document files (DOC), PDF, XLS, RTF or normal text documents. Table 2 illustrates the SOAP API which is used in the proposed framework.

Table 2: the soap api description

Service Function	**WSDL URL Location**	**Class**	Parameters
OCRWebService Recognize	www.ocrwebservice.com/services/OCRWebService.asmx?WSDL	OCRWS Response	User Name License OCRInputImage OCRWSSettings OCRWSResponse

The OCR REST approach is easy to use in developing requests. Accordingly, any browser can access the URLs, through HTTP client using any programming language to interact with the API. Figure 8 describes the OCR SOAP APIs and REST APIs parameters and related algorithms for each.

4.3. Layout Analysis and Performance Metrics

In this section, segmentation algorithms will be evaluated from practical argument of assessment. Such algorithms functioning after preprocessing phase, i.e. the image is the result after binarization module. Because, most OCR systems required binary images based on connected components. The regions we are dealing with, contain only text content. So, wide range of languages and documents categories will be handled. Finally, some algorithms require training dataset, they may have some attributes for training phase. Accordingly, the evaluation of the segmentation algorithms will be based on functionality of the algorithms. The current paper focuses on the developments, strengths and weaknesses of the OCR segmentation algorithms.

Therefore, performance evaluation of the scoring for each algorithm will compute the precision and recall for each attribute. The most top published dataset languages are English, German, French, Arabic and Chinese.

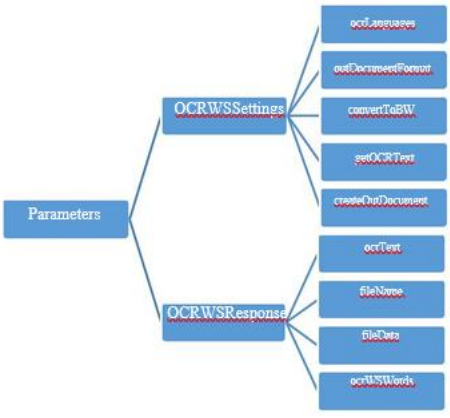

Fig. 8: The SOAP and REST parameters classification and related Tasks

The two main questions of this paper are: (1) What are the different dataset of OCR system will be covered?, and "What are the documents categories types can be covered?" (2) What are the functions of the segmentation algorithms will be evaluated

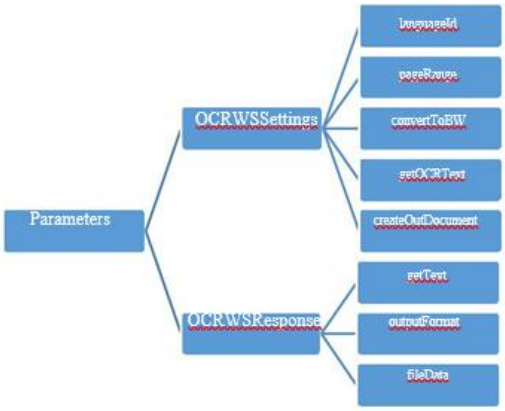

Fig. 9: The SOAP and REST parameters classification and related Tasks

Table 3 summarizes the main algorithms which are used in segmentation phase in OCR systems, take into consideration the type/kind of the processed documents.

Table 3: the segmentation algorithms with related functions

Algorithm	Layout	Color Depth	Labels	Training	Output type	Text Orientation	Text alignment	No of Language	No
Segmentation[12]	Any	BW	Yes	Yes	Text lines	Horizontal	Straight	3	6
Segmentation[13]	Any	Color/BW	Yes	No	Text lines	Horizontal	Straight	1	2
Line Extraction[14]	Any	Color/Gray	Yes/No	Yes	Regions	Any	Curved	1	2

5. Testing and evaluation

To evaluate the proposed framework, using different categories of dataset (Arabic calligraphy documents, and Arabic handwritten documents). Such experimental test validates each individual module in the preprocessing phase, as well as for testing the accuracy of the OCR system. We need 2 sub-categories:

(A) 20 images gathered from Arabic calligraphy documents.
(B) 20 selected images taken from Arabic typewritten documents.

Datasets of any imaged-documents are very important part to measure the accuracy of system. The accuracy achievement is very important to measure the performance of the recovery model. The percentage value of the accuracy of the proposed OCR system for any documented text can be calculated by the following equation:

N: represents the total number of words in the reference file.
D: stands for the deleted words in the resulted file.
S: represents the substituted words in the resulted file.
I: is the inserted words in the resulted file.

The resulted images form de-noising process were tested using three engines ABBY, Sakhr, and the proposed Arabic OCR system (See Table 4).

Table 4: average recognition results

	OCR Recognition Accuracy		
Pages type	***ABBY Engine***	***Sakhr***	***Proposed OCR***
Noisy pages	35.34%	23.21%	**41.12%**

OCR Recognition Accuracy			
Pages type	***ABBY Engine***	***Sakhr***	***Proposed OCR***
Sparse denoising	56.93%	37.66%	**60.37%**

The proposed line segmentation algorithm takes its input as calligraphy, typewritten and/or handwritten imaged-documents, and produces segmented text waved (straight or curved) line images. The proposed algorithm is tested on 40 different imaged-documents with 1480 lines (35-40 text lines for each image). Table 5 summarizes detail results of lines segmentation.

Table 5: the segmentation accuracy of the calligraphy books

Book	No. of segmented lines	Accuracy
1	37	99.94
2	35	99.95
3	38	99.96
..	..	..
39	35	99.93
40	36	99.92
Total	1480	99.94 %

The line segmentation approach that we developed in this paper is a combination between the connected components based approach and the seam carving based approach. This integrated approach takes advantage of the seam carving to segment noisy documents and makes better linking of the dots and the other Arabic diacritics. Table 3 shows the line segmentation results for sample pages of Arabic calligraphy, and old documents of our data. Also, table 5 shows the impact of this developed line segmentation on the OCR recognition results. The results in table (6) shows that the OCR new version of the system provides absolute 62% absolute gain over the previous version.

Table 6: average line precision

Algorithm type	OCR (Release1)	OCR (Release2)	Gain
	Line precision	Line precision	
Seam carving	73.08%	89.11%	16.03%
Proposed algorithm	72.79%	95.46%	22.67%

Table 7 summarized the percent accuracy of the recognition process of the used calligraphy and typewritten books.

Table 7: accuracy percentage for calligraphy books

Book	Accuracy	Book	Accuracy	Book	Accuracy	Book	Accuracy
1	82%	11	78%	21	68%	31	67%
2	79%	12	77%	22	70%	32	67%
3	78%	13	76%	23	70%	33	65%
4	80%	14	74%	24	70%	34	63%
5	78%	15	76%	25	64%	35	65%
6	79%	16	74%	26	64%	36	66%
7	79%	17	70%	27	69%	37	65%
8	76%	18	73%	28	60%	38	65%
9	75%	19	74%	29	68%	39	78%
10	71%	20	72%	30	67%	40	81%
Average				73.00 %			

6. Conclusions

This paper discussed a number of the major areas where OCR web services have been working comprehensively with supporting of document analysis and OCR service-oriented architecture. In this paper we targeted to select the suitable OCR algorithms to enhance the practical accuracy over the whole processes of the layout analysis, segmentation and recognition. In this work we worked on enhancing the accuracy of all the components of OCR system starting with document preprocessing and denoising, text detection, line segmentation, features extraction and the recognition processes. Some new techniques are introduced and other are enhanced. Also we addressed the challenges of old Arabic documents and managed to improve the OCR accuracy of such type of documents above the level to make them searchable documents. The system used the modified HMM classifier to extract features and therefore recognize the input documents into text. The suitable OCR system is tested and achieved accuracy not less than 73 % for calligraphy, typewritten and handwritten.

Acknowledgment

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH) – King Abdulaziz City for Science and Technology -the Kingdom of Saudi Arabia– award number (11-INF-1997-03). The authors also, thanks Science and Technology Unit, King Abdulaziz University for technical support.

References

[1] S. Setlur and Z. Shi, (2014). "Asian character Recognition", D. Dormann, K. Tombre (Eds.), Handbook of Document Image processing and Recognition, DOI 10.1007/978-0-85729-859-1_14, Springer-Verlang London, pp. 459-486.

[2] H. Cao and P. Natarajan, (2014). "Machine printed character recognition", D. Dormann, K. Tombre (Eds.), Handbook of Document Image processing and Recognition, DOI 10.1007/978-0-85729-859-1_44, Springer-Verlang London, pp. 331-358.

[3] H.Al-Barhamtoshy, and M. Rashwan, (2014). "Arabic OCR Segmented-based System", Life Science Journal, 11 (10), (ISSN: 1097-8135),http://www.lifesciencesite.com/lsj/life1110/200_27304life111014_1273_1283. pdf&sa=X&scisig=AAGBfm0YM6ykkOm8jGglYVhx2mT-ZU8OIA&oi=scholaralrt, http://www.lifesciencesite.com.

[4] U. Pal, and N. Dash, (2014). "Language, Script, and Font Recognition", D. Dormann, K. Tombre (Eds.), Handbook of Document Image processing and Recognition, DOI 10.1007/978-0-85729-859-1_9, Springer-Verlang London, pp. 291-330.

[5] S. Zha, X. Peng, H. Cao, X. Zhuang, P. Natarajan, and P. Natarajan, (2014). "Text Classification via iVector Based Feature Representation". 11th IAPR International Workshop on Document Analysis System, IEE, pp. 151-155.

[6] K. El-Gajoui and F. Ataa-Allah, (2014). “Optical character recognition for multilingual documents”: Amazigh-French Abstract-Optical, IEEE Second World Conference on Complex Systems, pp. 978-1-4799-4647-1.
[7] M. S. Khorsheed and H. Al-Omari, (2011). “Recognizing Cursive Arabic Text: Using statistical features and interconnected mono-HMMs”, 4th IEEE International Congress on Image and Signal Processing, pp. 1540-1543.
[8] Krayem, N. Sherkat, L. Evett, and T. Osman, (2013). “Holistic Arabic Whole Word Recognition using HMM and Block-based DCT”. 12th International Conference on Document Analysis and Recognition, pp. 1120-1124.
[9] M. Baechler, M. Liwicki, R. Ingold, “Text line extraction using DMLP classifiers for historical manuscripts”, in: Proceedings of 12th ICDAR, IEEE, 2013, p. 1029.
[10] S. Cholia, D. Skinner, and J. Boverhof, “NEWT: A RESTful service for building High Performance Computing web applications,” in 2010
[11] Gateway Computing Environments Workshop, 2010.
[12] Lamiroy and D. Lopresti, “An Open Architecture for End-to-End Document Analysis Benchmarking,” in 2011 International Conference on Document Analysis and Recognition, sep 2011, pp. 42–47.
[13] H. M. Al-barhamtoshy, (2016). “Towards Large Scale Image Similarity Discovery Model”, 2nd International Conference on Advanced Technologies for Signal& Image Processing ATSIP’2016, March 21-24, Monastir Tunisia, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7523047
[14] S. Eskenazi, P. Kramer, J. Ogier, (2017). “A Comprehensive Survey of mostly Textual Document Segmentation Algorithms”, since 2008, Pattern Recognition 64 (2017) 1-14.