Journal of Engineering Technology
Volume 6, Issue 1, Jan. 2018, PP. 479-495

RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

Arabic OCR Metrics-based Evaluation Model

Hassanin M. Al-Barhamtoshy [1] and Sherif Abdou[2]

[1]Computing and Information Technology, King Abdulaziz University (KAU), Saudi Arabia
[2]Faculty of Computers and Information, Cairo University, Cairo

Abstract: The benchmarking of any technology is an important process to evaluate its performance compared with state of art progress for that technology. The benchmarking of OCR systems is a complex process since it relies on the performance of several modules that usually construct a processing pipeline that starts with image preprocessing, followed by layout analysis for the processed page, then line and word segmentation for the text blocks and finally comes the recognition step. The evaluation process of the whole OCR system should consider the evaluation of each one of these subsystems. In this paper, we survey the proposed approaches for benchmarking end-to-end OCR systems. We also introduce an application of the evaluation process to an Arabic OCR system with available tools that influence the indicator metrics used today. In addition, the specification and preparation process of the benchmarking data set is also discussed since it has great influence on the evaluation results.

Keywords: Arabic OCR, segmentation, recognition, accuracy, performance evaluation.

1. Introduction and Literature Review

With the recent dominance of using computerized systems in companies, enterprises and organizations, the work documents are produced originally in electronic formats. However, usually the official versions of the documents need to be signed or stamped and are saved and used as a scanned image. These image-based documents need to be converted to text-based documents to allow for any type of information retrieval and processing from these documents. Optical Character Recognition (OCR) is the technology used to convert scanned images of typewritten, handwritten or printed text into machine-encoded text. Besides the daily need for OCR tools, there is the unequaled heritage of Arabic content in the whole Arabic region. There are millions of Arabic books, magazines, newspapers and other types of documents archived in libraries and archiving organizations. Recently, several large-scale digitization projects, [1-4, 16], have managed to transform several millions of pages from Arabic printed heritage into digitally available resources with the aim to fully integrate the Arabic intellectual content into the modern information. Most of these digitized documents include some meta data that register some basic information about the documents and some of them include complete text transcripts that are generated automatically by OCR.

Compared with Latin based OCR tools, the currently available Arabic OCR tools are still lagging in performance with significant drop in accuracy [1]. This can be due to several challenges in the Arabic script, such as graphemes connectivity, dotting, multiple graphemes for the same character at different positions, and composite ligatures. The performance of Arabic OCR tools is even more deteriorated for historical documents, with low page quality and complex layouts, which make the produced text useless for retrieving information.

Several research efforts, in either academia or industry, are working towards improving the performance of Arabic OCR technologies. To have an accurate measure for the state of art performance of Arabic OCR systems, we need an effective metrics that can provide fair and accurate detailed evaluation for these systems. Several evaluation metrics have been produced for evaluating Arabic OCR systems, but most of them have focused on specific evaluation for part of the full OCR system. Actually, the mostly referenced metric is the Word Error Rate (WER) that focuses on evaluating the recognition accuracy of OCR systems. The full OCR systems consist of several processing operations, starting from binarization, noise removal, de-skewing, line and word segmentation, then finally the recognition process. The performance of each processing step affects the performance of all the following steps. Therefore, we need to have an effective metric for each one of these steps, in order to evaluate accurately the whole OCR system performance.

Besides the evaluation metrics, there are also the evaluation data sets. The evaluation test set has to be generalized enough to provide accurate measure for the real life performance. For example, the system performance for different fonts, sizes and styles. In addition, whether the system can process old historical documents with similar performance for modern documents. Also, the capability of the system to recognize camera captured documents or the closed caption text embedded in video recordings. All these issues need to be considered when constructing a benchmark OCR evaluation test set and even when constructing the main training data set for the system.

Creation and annotation dataset for document analysis and recognition was initiated in a benchmarking of the OCR purpose based with evaluating well-known metric presented in [1, 2]. Therefore, an algorithm was implemented in an OCR system for annotating one of standard handwritten Farsi/ Arabic digit datasets [3]. An Arabic evaluation tool is presented in [4], with accuracy metric and an enormous number of testing experimentations. Consequently, word-based and character accuracy metrics are presented to evaluate the performance of Arabic OCR systems [5]. Three types of scanned images are used (newspapers, books, and journals) [5]. The trial results showed that such metrics are valuable for evaluation of the Arabic OCR systems.

An annotated video dataset is implemented to evaluate Arabic text detection and tracking in video stream. The video dataset is collected from four different Arabic news channels. It consists from 80 videos with 850,000 frames of maximum diversity of content sizes [6].

Before developing metrics-based evaluation model, some definitions are needed to be used in this literature.

- **Evaluation.** American Evaluation Association defines evaluation as "assessing the strengths and weaknesses of programs, policies, personnel, products and organizations to improve their effectiveness' " [http://www.eval.org/].
- **Ground Truth (GT).** This term refers to the dataset that is collected "on location". In our case, the term of "ground truth" means the classification of the dataset, used for training and testing the proposed systems.
- **Benchmarking**. As defined in Oxford Dictionary, means "a standard or point of reference against which things may be compared".

Beside the previous three definitions, there is the collection of real dataset with respect to the collection of segmented and recognized text data. Arabic document images scanning, labeling and annotation are not easy tasks, the cost of these tasks is high during preparation of the diversity of Arabic document images [7]. Very important questions that may be asked at this level; are:

1. What are the domains that should be covered in Arabic document datasets needed to start with?
2. What is the planned function and tool of labeling task during the training phase?
3. What is the size of the Arabic document pages to be processed at training and testing phases?

4. How to measure the accuracy? What are the costs of different metrics made by the proposed system?

The remainder of the paper is systematized as the following. Section 2 introduces to the “Proposed evaluation metrics” to assess the Arabic OCR systems with existing datasets. Therefore, the implementation overview within related criteria will be described in Section 3. Implementation overview and experimental validation with results will be pronounced in Sections 4 and 5. Section 6 includes concluded remarks and future work prospective.

2. Methodology of the Proposed Evaluation Metrics

Document quality plays very important task in image analysis and recognition tasks [3, 7]. So, document quality metric is used to measure document segment and recognition qualities. The following sections describe a group of evaluation methods to be used in our solution.

To design and implement a methodology of benchmarking in OCR systems, the modules of this system are considered for evaluation in general. Considering Arabic OCR system as a series of modules, the detection of detailed view is needed. As a result, what is required to be measured may be a single module, a series of components or a subsystem of the whole system, see Fig. 1.

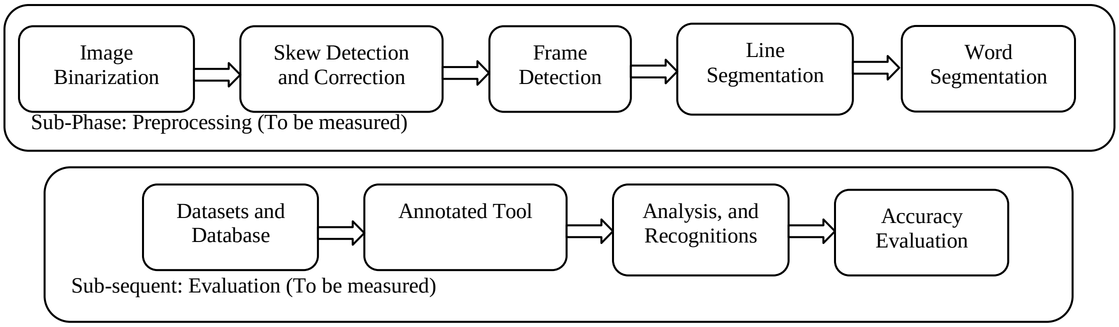


Image Binarization
Skew Detection and Correction
Frame Detection
Line Segmentation
Word Segmentation
Sub-Phase: Preprocessing (To be measured)
Datasets and Database
Annotated Tool
Analysis, and Recognitions
Accuracy Evaluation
Sub-sequent: Evaluation (To be measured)

Fig. 1. System in OCR metrics

Proposed benchmarking systems had been carried out for years before now. In general established tests are carried out by institutions like NIST and SRI [2]. In all the phases of OCR system, the need of benchmarks has been declared, other than those few that have been published since then. This is primarily due to the reality to result the best measurement is difficult, and there is no standard describing experiment dataset with ground truth. These problems formulate benchmarking much harder in preprocessing than in other phases, such as classification and recognition. Therefore, different details of intermediate analysis results, final system output results, and the output will be compared with ground-truth imaged-documents. Such intermediate analysis includes binarization, skewing, noise removing, segmentation, training, and recognition. Such previous modules would be applicable for imaged-documents analysis as the implemented algorithms. Any OCR experimental evaluation considers a system as a black box, and the evaluation is based on the final system results (outputs) with specific aspect of quality. However, when creating different details, and complete accuracy assessment, this methodology cannot be adequate.

The proposed evaluation model includes the following: (1) Document datasets with different domains. (2) Annotation tool to describe the corresponding datasets and building the ground-truth that describes each document with its labeled information [1, 3, 11, and 14]. (3) Standard formats to store the document datasets (such as TIFF, JPEG … etc. for inputting images and XML to define the description of the ground-truth [3]). (4) Physical structure and datasets organization that permits easy access and manipulation of different categories of the datasets [1]. Fig. 2 explains general data flow diagram of the evaluation metrics of Arabic OCR system. The proposed process starts by scanning the Arabic imaged-document; different tasks of the OCR system take place (binarization, skew detection and correction, frame and noise detection and removing, segmentation [11], and recognition). Therefore, we get the output of the Arabic OCR and judge it through the ground-truth dataset. Finally, we visualize the performance of the system (Fig. 2).

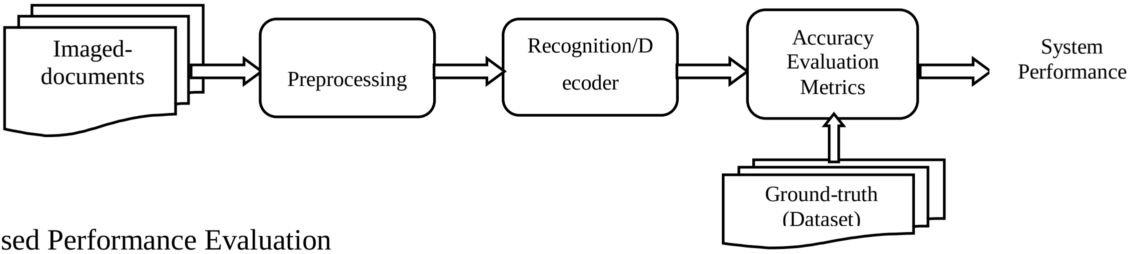


Fig. 2. Proposed Performance Evaluation

Accordingly, table 1 illustrates the proposed metrics as a benchmark to differentiate between different OCR systems.

TABLE 1 DEFINITIONS OF INTEGRATED BENCHMARK AS METRIC IN OCR'S SOFTWARE

Metric Items		Description
1	OCR Name	The name of the OCR software as a product.
2	Accuracy	The degree to which an OCR measurement result conforms to the correct value or a standard value of that measurement.
3	Performance	The achievement of a specified OCR's task measured against the standards of accuracy, completeness, cost, and speed. How well an OCR task does a piece of work, at any speed at which a computer operates.
4	Dataset	The collection of documents; related, or discrete in different categories and images of domains, which may be used in training and testing or as a whole.
5	Skew/De-Skew	To detect the skewing angle for each document/ segmented objects. The De-Skew is the rotation of the whole page/ segmented object by this skewing angle
6	Noise/De-Noise	The noise is defined as variation of brightness or color information in the documents and can be measured by degradation of the image by using connected component algorithm or other methods.
7	#Languages	The problem of identifying the language that indicates to the text regions content.
8	Input Format	Different input formats that can be used as reading the image (pdf, png, jpg, tiff, …)
9	Format	Different output format to create the recognized plain text (e.g. txt or doc).
10	Web Services	Any piece of OCR can be accessed by the internet and integrated with new web-based applications using open standard (XML).
11	Cloud-Storage	The computing virtual or real model in which dataset is stored on remote servers accessed by the web services.

2.1 Existing Datasets

The first category of the dataset used in binarization methods according to visual evaluation related to a set of predefined criteria [1]. Some generated datasets organized in international conferences, such as ICDAR [16], DIBCO [9, 17], and ICFHR [18, 19]. Such datasets consist of machine printed images and handwritten text, see table 2. In addition, text recognition dataset has been created to be used in evaluation criteria, table 2 includes comprehensive list for Arabic text recognition [20].

TABLE 2-A BINARIZED DATASETS

Dataset Name	No of Images
DIBCO'09 [3]	10

TABLE 1-B ARABIC TEXT RECOGNITION DATASETS

Dataset Name	No of Images
APTI	45,313,600 Words

H-DIBCO'10 [17]	10
H-DIBCO'11 [9]	16
ICFHR'10 [18]	60

IFN/ENIT	26,4000 City names
IBN-SINA	1,000 Sub words
ADAB	15,158 Words (Different writers)
CENPARMI	23,325 Handwritten Digits

In the field of miscellaneous document, analysis, text localization, recognition, and other datasets have been created in table 3, obtained from non-paper works.

TABLE 3 BINARIZED DATASETS WITH RELATED TASKS

Dataset Name	No of Images	Tasks to be used
ICDAR'11 [16](Challenge 1)	420	Text localization and Text recognition
ICDAR'11 [16] (Challenge 2)	485	Text localization and Text recognition
KAIST [19]	3,00	Text localization and Text recognition
Google street view [20] [21]	350	Word Spotting
IUPR [21]	100	Zone segmentation and Recognition
NEOCR [9]	659	Text localization and Text recognition

2.2 Creating Annotation Tool

One page document needs about one hour to be fully annotated [1] in the case of region segmentation [11]. Unfortunately, manual annotation requires at least two persons (one for creation and another for validation) for one page preparation in ground-truth. Consequently, many interactive tools are described in literature [15] that take into consideration page or document definition and validate region segmentation. Such region segmentation is presented in different shapes; i.e. rectangles, polygons, or isothetics. Additional information, such as labeling, are associated to each segment; content types or relation with other segment, and stored such information in standard format like XML. The ground-truth simply is the process of labeling imaged-characters, imaged-words; or imaged-lines. Moreover, the labeling or the process of text transcription is used as a basis of validation and correction. The automation of the proposed annotation tool needs three strategic methods. The first strategic process is the type of application domain of the dataset (e.g. printed book, typewritten book, historical book, handwritten text, manuscripts …etc.). The second strategic process is the determination of the basic operations of the Arabic OCR system (preprocessing, analysis, feature extraction and recognition). The third strategy is verification and validation of the recognition process with respect to the ground-truth dataset. Accordingly, three tasks are needed in the proposed annotation tool:

1. Document and page imaging that can be used in binarization, skewing, and noise removing.
2. Page analysis during segmentation, feature extraction and classification processes.
3. Text recognition for printed, calligraphy and handwritten documents inside the two cases, offline and online.

The proposed annotation tool allows adjusting selectable regions, paragraphs, lines, or words by freehand lines.

2.3 Ground-Truth Dataset Description

It is required to produce a large dataset for Arabic text to assist in advanced research and product prototyping of Arabic OCR. The database is required to consist mainly of documents (one page per image), and the equivalent formal explanation of that image (XML transcript file). Therefore, an attempt to solve analysis of the physical structure of the document, and analysis of homogeneous components are presented [6]. Such paper introduces to "Universal Datasets Repository for Document Analysis and Recognition (UMDAR)". UMDAR helps dataset creators to standardize their datasets and making them accessible by the research community once published on the proposed repository [6].

The number of scanned documents produced is anticipated in order of 15,000 images, with 15,000 corresponding transcription files. These are mainly using two streams; the first is generated using word lists and the second is generated using a collection of books, journals, thesis, calligraphy and typewritten documents. The collection came from 15 different categories, based on the used fonts and sizes. The used collection is classified manually and approved by human experts. Therefore, the selected number of pages is around 15000 images during 10 years (at least average of 10 pages from each category for copyright constraints, which gives approximately 150 imaged-documents). The domains of the categories are chosen to cover uniformly the past 50 years. Consequently, 1000 pages from thesis (in Arabic) have been selected as well, which also covered uniformly the past years. Consequently, this dataset can be annotated by a supervised and/or semi-supervised by the designed proposed tool.

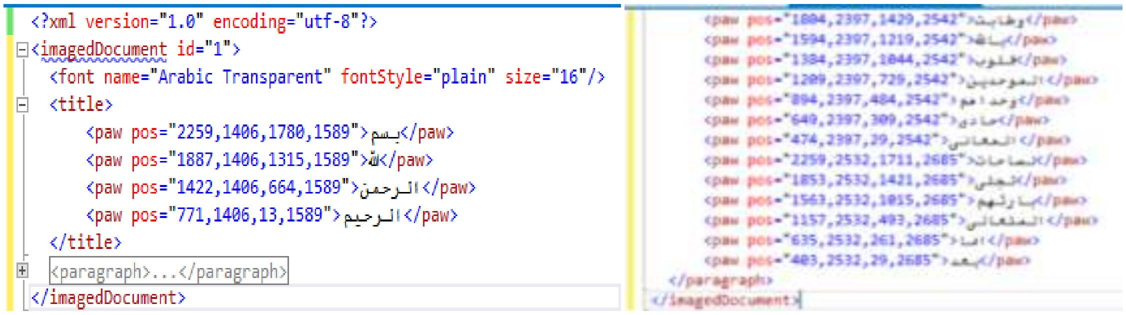
2.4 Fonts and Sizes for the Word Lists

As we mentioned in UMDAR dataset [7], it included a universal datasets repository for document analysis and recognition (UMDAR). Dataset creators need to standardize their datasets and making them accessible by the research community once and stack holders. In addition, it can be used as a central, which bridges, in a smart manner, between datasets description and all document analysis stakeholders.

The generated datasets of the strategic requirement will be carried out according to the following specifications:

- Fonts include: (1) Simplified Arabic, (2) Arabic Transparent, and (3) Traditional Arabic.
- Sizes: each of the above fonts is required to be produced for those sizes (except the typewriter and calligraphy): 10, 12, 14, 16, 18, 20, and 22.

Each segmented line, word or sub-word (Piece of Arabic Word: PAW) in the Arabic OCR dataset is fully described using XML standard format. The labeling of the segmented line, word, or sub-word contains tags that represent ground truth information about sequence of words/PAWs. Fig 3 gives an example about generated information of imaged-document, line text with PAW tag as segmented locations.



```
<?xml version="1.0" encoding="utf-8"?>
<imagedDocument id="1">
  <font name="Arabic Transparent" fontStyle="plain" size="16"/>
  <title>
      <paw pos="2259,1406,1780,1589">بسم</paw>
      <paw pos="1887,1406,1315,1589">ﷲ</paw>
      <paw pos="1422,1406,664,1589">الرحمن</paw>
      <paw pos="771,1406,13,1589">الرحيم</paw>
  </title>
  <paragraph>...</paragraph>
</imagedDocument>
```

```
      <paw pos="1804,2397,1429,2542">وطابت</paw>
      <paw pos="1594,2397,1219,2542">بـه</paw>
      <paw pos="1384,2397,1044,2542">قلوب</paw>
      <paw pos="1209,2397,729,2542">الموحدين</paw>
      <paw pos="894,2397,484,2542">وحدا هم</paw>
      <paw pos="649,2397,309,2542">ما دي</paw>
      <paw pos="474,2397,29,2542">المعاني</paw>
      <paw pos="2259,2532,1711,2685">نصا مات</paw>
      <paw pos="1853,2532,1421,2685">تجلي</paw>
      <paw pos="1563,2532,1015,2685">با رئهم</paw>
      <paw pos="1157,2532,493,2685">المتعالي</paw>
      <paw pos="635,2532,261,2685">ما</paw>
      <paw pos="403,2532,29,2685">بعد</paw>
  </paragraph>
</imagedDocument>
```

Fig. 3 Example of XML description for ground truth information about given imaged-document

3. Proposed Metrics in OCR System

The quality of the images' dataset plays an important task in image analysis and recognition tasks [3]. Accordingly, document quality metric will be involved to measure document segment and recognition qualities. The following sections describe a group of evaluation methods to be used in our solution.

Two adopted metrics will be used: structural and statistical. For structural metrics, category of the images dataset, type and domain of the imaged-documents, and the related tasks at the OCR system. So, the structural methods use information labeled of the imaged-documents data to train classifiers for word/sub-word recognition purposes. For the statistical metrics; individual indicators inside each modules and OCR methods, such as binarization level, noise removing, skewing adoption, feature extraction and recognition metrics. Such metrics are defined locally at the method level.

3.1 Peak Signal-to-Noise Ratio (PSNR)

One of the very important evaluation methods is the "peak signal to noise ratio" (PSNR). This method can be calculated by the ratio of the "maximum power of a signal image", and the "power of noise affecting the quality of signal image". Accordingly, the "mean square error" (MSE) is used to calculate the PSNR, as the following two formulas.

$$\text{MSE} = \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} \frac{(I_{1\,(x,\ y)}\ x\ I_{2\,(x,\ y)})}{N_x\, N_y} \qquad (1)$$

Thus;

$$\text{PSNR} = 10\ .\ Log_{10}\ (\frac{Max(I)^{\ 2}}{MSE}) \qquad (2)$$

Where $Max(I)$ represents the maximum difference between foreground and background intensities of pixels. MSE is the mean square error. Therefore, PSNR calculates the difference of the document under testing against the ground-truth binary document, as shown in equation 2. Such two equations are used in dataset preparation, information retrieval and information extraction.

3.2 Recall, Precision and F-Measure (FM)

The recall and precision metrics are two measurements used to compute F-measure (FM) in many researches [5-9], and they are used in feature extraction; like speech recognition, OCR [8] and information retrieval domains [5],[9]. The recall and precision can be calculated according to formulas, shown in the following (Equation 3 and Equation 4).

$$\text{Recall} = \frac{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} GT(x\ ,\ y)\ x\ B(\ x\ ,y)}{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} GT\ (\ x\ ,y\ )} \qquad (3)$$

$$\text{Precision} = \frac{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} GT(x\ ,\ y)\ x\ B(\ x\ ,y)}{\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} B\ \ (\ x\ ,y\ )} \qquad (4)$$

Where GT represents ground-truth documents (original images), and B denotes binary test document image. As mentioned in other literatures [3, 8], the F-measure can be calculated according to equation 5.

$$FM = \frac{2\ x\ Recall\ x\ Precision}{Recall+Precision} \qquad (5)$$

It is easy to conclude that, FM is the ratio of ground-truth document and the corresponding same pixels in the binary document image under test to the ground-truth binary document image in case recall and the binary document under test in precision case [3].

3.3 Negative Rate Metric (NRM)

The negative rate metric (NRM) is the relation between the original document; ground-truth (GT) and the binary document (B) pixels. Four terminological values are needed to find the combination between the GT and the B documents, as shown in table 4.

TABLE. 4. THE RELATION BETWEEN GROUND-TRUTH AND BINARY IMAGE

	Condition Positive	Condition Negative
Ground Truth	TP	FP
Binary Image	FN	TN

Where; TP represents true positive; FP represents false positive; FN represents false negative. TN represents true negative. Accordingly, NRM is computed using equation 6, as follows:

$$NRM = \frac{Recall + Precision}{2} = \frac{\frac{FN}{FN+TP} + \frac{FP}{FP+TN}}{2} \qquad (6)$$

The recall and precision metrics are two measurements used to compute F-measure (FM) in OCR system.

3.4 Documents Analysis

This analysis includes segmentation and the layout or physical structure analysis of the imaged-document. Domain dataset analysis is an important key in document and image understanding. A summary of domains' number corresponding state of the art is given in table 5. The dataset includes logical labels that can be extracted as features across these domains.

TABLE 5. A SUMMARY OF DOMAINS' NUMBER CORRESPONDING STATE OF THE ART

Domain	Number of Arabic Documents	Syntactic Labels (Fonts, Sizes)
New Books	15	All fonts and all sizes
Journal	10	All fonts and all sizes
Thesis	10	All fonts and all sizes
Typewritten	10	Standard fonts and sizes
Calligraphy	10	Good Writers

The evaluation metric is based on ICDAR 2005 (page segmentation computation [3]), by counting number of metrics between the analyzed image and the entities of the ground-truth image, using global Match Score.

In the dataset preparation stage, three operations are used; scanning, data annotation and data revision of the input images, before the evaluation metric takes place. The second phase is the "training phase", which includes a preprocessing techniques (such as binarization, skewing, filtering, etc.), features extraction, verification model to look at and revise the segmented data (by human expert and additional tools). Then, the "evaluation metric" of the training phase takes place to evaluate the output results of this phase. Entirely, there is an enhancement phase in order to generate very accurate dataset. In addition, in the training phase, word generation features extraction and clustering process will be employed. Finally, the "recognition phase" is used to recognize the estimated document image, clean lines, words and characters. A detailed description of the features extraction module can be found in [11, 12]. The default concept is based on the property of the discrete cosine transform (DCT) in the "recognition phase". In addition, a transformation coefficient has been selected as a word feature to generate a unique vector and minimize errors on the classification process.

3.5 Metrics Evaluation

As mentioned in [7], [8] and [11], the data preparation is important during document understanding, document analysis and the OCR processes. In these processes, data may contain unusable image formats, missing values, errors, and compressed format. Therefore, additional tool may be used at this level. Moreover, segmentation of page(s) into logical elements is very important to be used in the evaluation as metric indicators. Figure 6 illustrates examples of document pages with related object elements. The page layout depends on, language identification [14], document orientation (portrait or landscape) and document type (book, journal, thesis, typewritten or calligraphy). The ground truth of the document assigns each physical block to a logical element of the tested image (text, image, or graphic).

ICDAR presented an evaluation method to be used in page segmentation [3, 7]. This method depends on the number of matches between the elements detected by the method and the elements in the ground truth (GT) document. A global “Match Score” method will be used to build match score table, from the previous results [3]. The values of such table are computed according to matching the ground truth and the image results.

Suppose that GT_i is the set of all segments inside the ground truth region i, R_j represents the set of all segments inside the result region j, I represents whole image segments, and T() counts the elements of the whole set of the document. Now, the *MatchScore*(i , j) is computed according to the following formula [1], (equation 7).

$$MatchScore(\,i\,,j\,) = \frac{T(GT_i \cap R_j \cap I\,)}{T(GT_i \cup R_j \cap I\,)} \qquad (7)$$

3.6 Document Binarization

As cited in many definitions that "the conversion of a color image or grayscale into binary image" [3], [6], [9, 10] is what we call image binarization. The binarization approach can be categorized into global, local, and hybrid methods. The global method uses single threshold value T for the entire document, so, the resulting binary document B(x , y) is defined as the following (equation 8):

$$\text{B (x , y)} = \begin{Bmatrix} 1, if\ D(x\,,y) \le T \\ 0, if\ D(x\,,y) > T \end{Bmatrix} \qquad (8)$$

However, image or document enhancement and therefore, noises is reduced at the preprocessing phase of the image processing and document analysis. The binarization techniques are based on several metrics, and they will be described in many literatures [3], [11]. Such techniques are basically depending on comparative methods between the binary documents and the corresponding ground-truth documents. Figure 5 shows an example of Arabic printed page as a document image, together with the corresponding original ground-truth binary image.

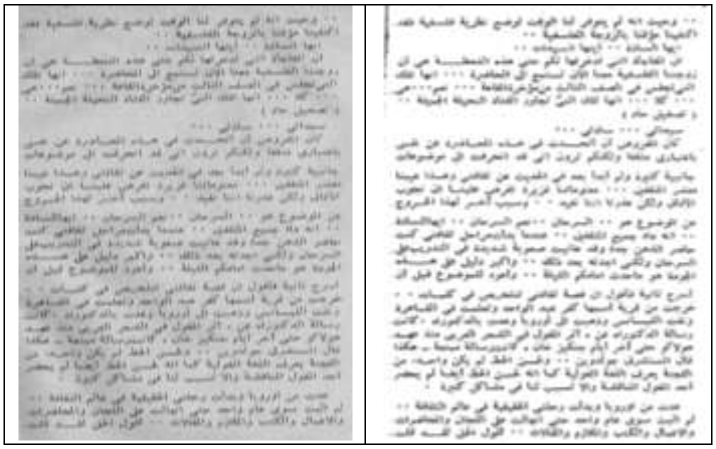
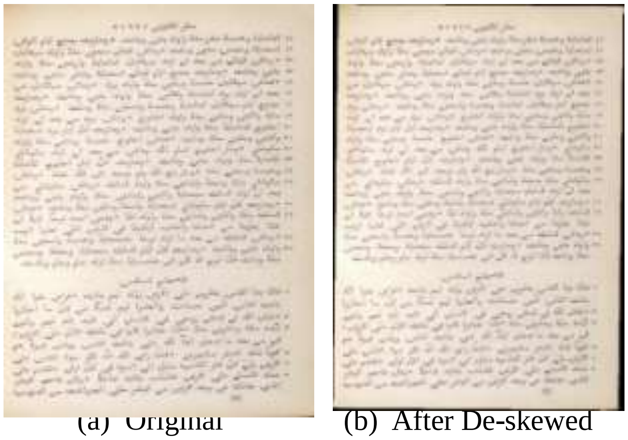


Fig. 5. Document of a part of (a) book page, and (b) the corresponding ground-truth (GT) document

3.7 Skew Detection and Correction

Our de-skewing algorithm is described as the following. First segment the image, then detect the skewing angle, but when we try to obtain the skew angle from page borders, we faced the problem that some pages do not contain any borders and some borders are not continuous, so it will be hard to be extracted. That leads us to try another technique in skew detection. All borders and small components are removed, then the page is segmented into lines using a histogram technique, curve fitting is then used to obtain the skew angles for each line, and finally the average skew angle calculated and rotate the whole page by this skew angle. The main advantage of this method is its ability to detect the skew angle, even if it is very small. Fig. 6 shows sample skew and de-skewed image.



(a) Original (b) After De-skewed

Fig. 6. Skewed and De-skewed images

3.8 Shadow and Frame Detection and Removing

Some Arabic calligraphy and Arabic historical documents are different from those in other categories. The difference is that the existence 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 7 shows the block diagram of the noise removal module. Figure 8 (a and b) shows sample processed image after noise removal.

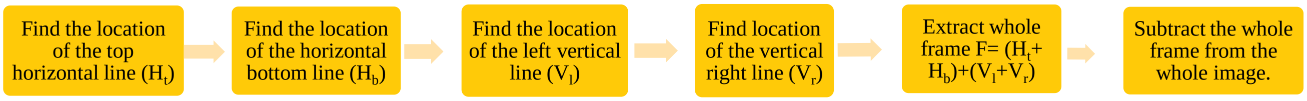
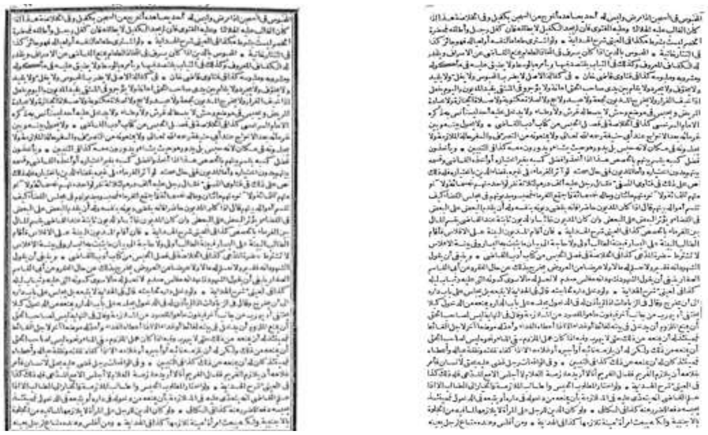

Find the location of the top horizontal line (H_t)
Find the location of the horizontal bottom line (H_b)
Find the location of the left vertical line (V_l)
Find location of the vertical right line (V_r)
Extract whole frame F= (H_t+ H_b)+(V_l+V_r)
Subtract the whole frame from the whole image.

Fig. 7. Frame detection and removing

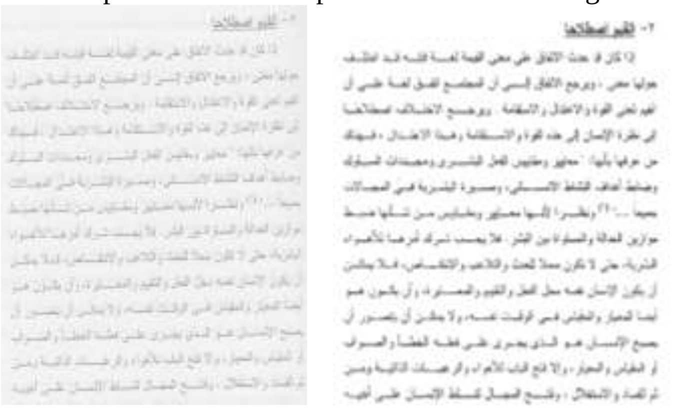
In addition, the proposed OCR uses the sparse algorithm to remove shadow noise that is a common effect in old and historical documents. The main advantages of this algorithm are its ability to use a learned prior knowledge in noise removal process. Therefore, to remove shadow noise, that is a common effect in old and historical documents, we use sparse representation [8]. The main advantages of this algorithm are its ability to use a learned prior knowledge in noise removal process.



(a) Image before frame removing (b) Image after frame removing

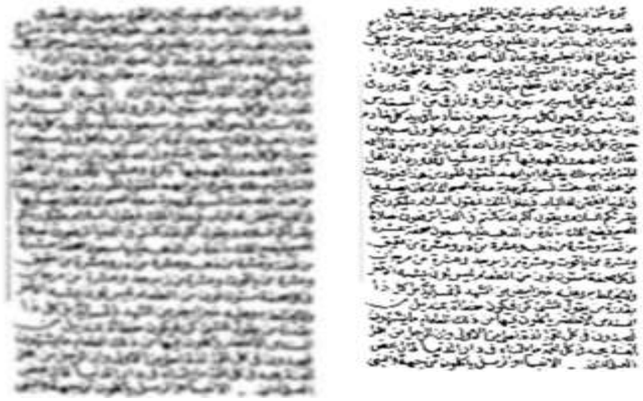
Fig. **8**. Frame detection and removing for Arabic document

These in addition to, some types of documents need to be enhanced using de-noising process. These types are calligraphy and historical documents. Therefore, removing images noise as well as preserving documents details and structures are important. Accordingly, total variation modified model with Rudin-Osher-Fatemi (ROF) smother process is used to preserve documents edges and structures (see Fig. 9).



(a) Before noise removing (b) After noise removing

Fig. **9**. De-noising using sparse representation



(c) Before noise removing (d) After noise removing

Fig. **9**. De-noising using ROF smoothing proces

3.9 Document Segmentation

As mentioned before in [11], text regions are segmented into lines, and each line is segmented into words' regions. Evaluation metric of the "text segmentation results" is compared to its corresponding label ground truth (GT), according to the following definition.

Definition 1. Let GT_i represents the i^{th} ground truth image, where i ϵ {1 … N_g}, and N_g is the number of text-lines in the ground image. Let T_j represents the j^{th} segmented text-line, where j ϵ { 1… N_l}, and N_l is the number of extracted lines. Then compare GT_i with T_j.

Again, as presented in "F-Measure" metric (Eqns. 3, 4, and 5), it is important to *MatchScore* (i , j) between GT_i and T_j. Accordingly, the algorithm decides if *MatchScore* is greater than or equal a threshold value T; (*MatchScore* >= T). Then, calculate the precision value in case of calculating detection rate (DR), as the following formula:

$$Precision = \frac{MatchScore\ (i\ ,j)}{N_g} \qquad (9)$$

In addition, recall computing can be calculated according to recognition accuracy (RA), as the following formula:

$$Recall = \frac{MatchScore\ (i\ ,j)}{N_r} \qquad (10)$$

The overall accuracy of the F-Measure is as illustrated in equation 5, as the following:

$$FMeasure = \frac{2\ x\ DR\ x\ AR}{DR + AR} = \ 2\ x\ \frac{\frac{MatchScore\ (i\ ,j)}{N_g}\ x\ \frac{MatchScore\ (i\ ,j)}{N_r}}{\frac{MatchScore\ (i\ ,j)}{N_g} + \frac{MatchScore\ (i\ ,j)}{N_r}}$$

$$= \frac{2\ MatchScore\ (i\ ,j)}{N_g + N_r} \qquad (11)$$

4. Implementation Overview

The overview of the proposed system is shown in Fig 4. It includes three main steps; first, the pixels in the given text/ paragraph region are grouped into cluster using K-means. Then, the output results (segmented text lines) will be detected and therefore, calculate the following actions: (1) Compute precision, (2) Compute recall, and (3) Compute F-Measure. Consequently, text cluster into segmented lines action are what we mean by segmentation.

(A) Clustering

K-means clustering will be used in the training and recognition phases, taken into consideration the binary images to group the pixels of the scanned image into K cluster [12]. During the training phase, features are extracted from the training set by performing algorithm to build feature vectors database for the training set using three font types, Simplified Arabic, Traditional Arabic, and Arabic Transparent with sizes 12, 14 and 16 respectively. In order to recognize a word image, the feature vector of word image is compared against each training feature vector by computing the Euclidian distance to measure the similarity between the two vectors. Then, prediction class of the testing image is found based on the minimum difference measured by the Euclidean distance; between the testing word image and the training samples.

In this work, the classifier identifies the top k closest vectors (classes), which have the minimum Euclidean distance with the test image vector, and then arranges the distances in ascending order representing the candidate words chosen as entries to the next stage. The Euclidean distance $d(w_c , w_t)$ for each candidate word w_c can be calculated as squared Euclidean distance or as absolute Euclidean distance.

(B) Verification Model

Verification model attempts to compute the overall metrics between the three phases of retrieving a word (W) given the F-Measure (FM) from the language model. Moreover, this model attempts to estimate similarity between words on a page as follow [13]:

$$Sim(W|F) = P(R|w) / P(!R|w)$$

Where R is the set of relevant words to a language model (features). This model tries to compute the probability that the word is relevant and the probability of the same word is not relevant [13].

5. Experimental Validation

This section presents the experimental results of the proposed Arabic OCR System that have been described. These results are based on the assumption that the test dataset will be performed by perfect word segmentation and line segmentation at the training phase and word recognition at the recognition phase. This type of dataset is composed of 50 pages taken from books (15), journals (10), thesis (10), typewritten (10) and calligraphy (10) of document images. Arabic documents that have been printed in three fonts and four different sizes. Figure 12 shows what it might look like to scan (left side), segment (right side), recognize and verification (right down side) for the scanned image using the proposed OCR system.

At the end of the recognition phase, the verification module, if there is difference between the number of lines in the segmented image and the recognized text, the verification module focuses on such differences by employing definition 2.

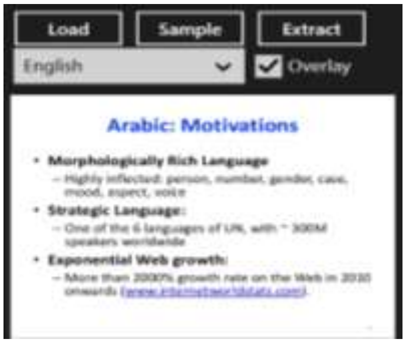
Definition 2. A document description D_d is a triplet D_d = { Language Id, ID ϵ Image, P_{size} C page size, F C set of features (pages $_i$, lines $_j$, words $_{I,j}$) }; where i ϵ {1 … Height and j ϵ {1 … Width }

Certainly, more identifying features could be added, such as paragraph number, document number, and so on. The advantage of having additional information used to identify words is the information that can be used later to define features for the machine learning algorithms.

(A) Overall Evaluation

As mentioned before, the widely used metrics in logical labeling are recall and precision. Accordingly, private ground-truth dataset is prepared for testing the state of the proposed Arabic OCR system, by providing images to evaluate the system. In this part, the proposed system has been tested against two other systems, using 15 document pages of dataset from diversity. Figure 10 shows and Arabic typewritten case testing for the Arabic OCR system.

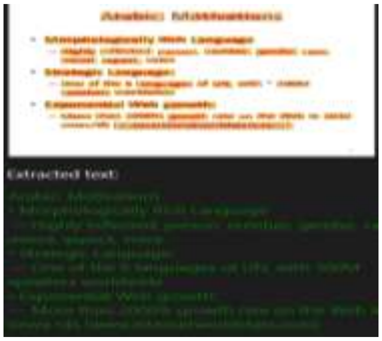
In order to analyze the calligraphy pages (good Arabic handwritten documents); let us briefly explain a process anticipated by Zheng et al. [22]. Such method employments connected components as elements of processing. Therefore, the goal here is to categorize connected components into either early printed or calligraphy/good handwriting. To complete this objective for noisy or not clear documents, Zheng et al. [22] describe their classification task into either machine printed, handwriting, or noise. Accordingly, the Zheng et al. task is improved in the three categories of the Arabic document types (early printed, books, and Calligraphy). The used extracted features are like those in page segmentation (e.g., Gabor filters). Next step includes "features extraction" from each connected component, with classification achievement. In other way, "Fisher classifier" is applied after feature selection by principal component analysis. Later the classification task is error prone due to the limited amount of information from each connected component, the language model and extra processing are necessary to filter the results by considering the appropriate and contextual information. Figure 11 (a, b, c, d, e, and f, and g) shows examples of the Arabic documents after binarization, noise removing and frame removing, skewing and de-skewing, and segmentation processes.



Arabic: Motivations

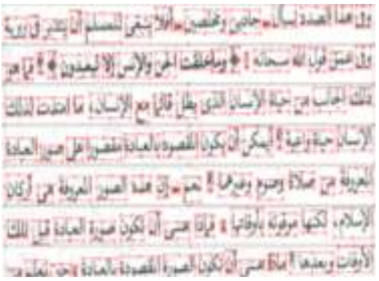
- Morphologically Rich Language
  - Highly inflected: person, number, gender, case, mood, aspect, voice
- Strategic Language:
  - One of the 6 languages of UN, with ~ 300M speakers worldwide
- Exponential Web growth:
  - More than 2000% growth rate on the Web in 2010 onwards (www.internetworldstats.com).

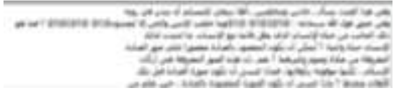
(a) English Scanned Document



Extracted text:

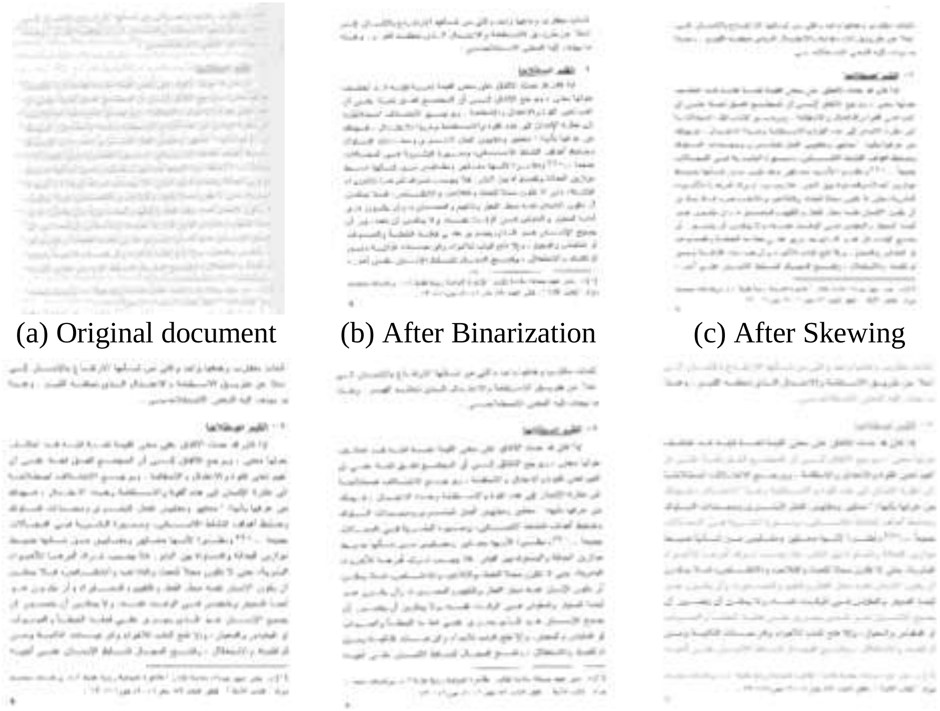
(b) Recognized English Text





(c) Arabic Scanned Document (d) Segmented & Recognized doc

Fig. 10. Examples of Scanning, Segmentation and Recognition Processes



(a) Original document (b) After Binarization (c) After Skewing

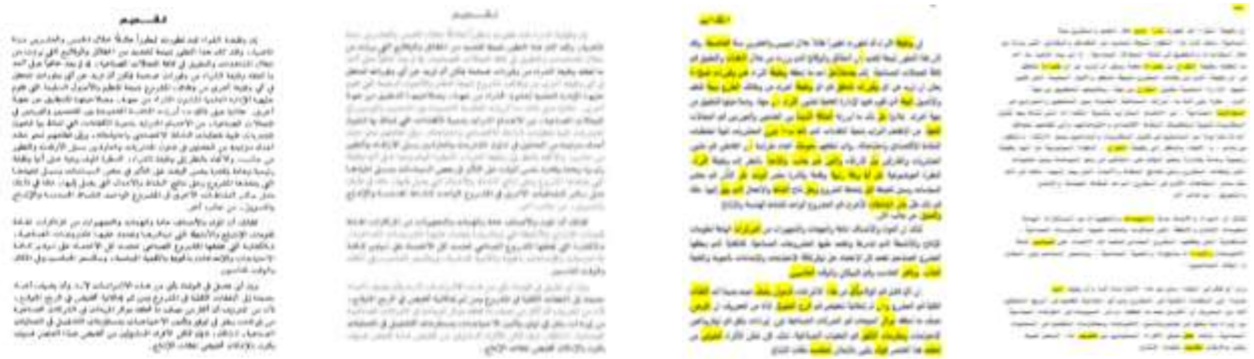
(d) After De-noising (f) After De-Framing (g) After Segmentation

Fig. 11. Example of Arabic calligraphy document after the pre-processing processes

Finally, the proposed system decides the value of the accuracy; such value is taking into consideration: (1) Correct segments for lines and words; (2) Wrong segments for lines and words (3) Correct characters/words; and (4) Wrong characters/words. Table 7 illustrates the evaluation results for the proposed system, taking into consideration human experts' judgment for the five collection dataset types. We compared the performance of our OCR system with the 2-D HMM based system as reported in, which is hybrid HMM-DNN with histogram features and use complete line recognition with 3-gram language model as descried in this report, and the two other systems. We used two test sets; the first one was pages selected randomly from modern Arabic documents. The second test set was pages selected from old, calligraphy and historical Arabic documents.

(B) Accuracy

The mentioned discussion is implemented and tested on Arabic document images with three categories of OCR system. Fig. 12 shows a sample of Arabic images (printed documents) that have been tested, taken into consideration three OCR systems. The comparative test has evaluated, taken into account the recognition accuracy of each system.



(a) Original image (b) Proposed Arabic OCR (c) ABBYY Fine Reader (d) Google OCR

Fig. 12. Original Image and the recognized text for three OCR Systems

(C) Speed

The comparative test has evaluated the five tasks (binarization, skewing/de-skewing, noise removing/ De-noising, frame removing/ De-framing, and segmentation) processes. In segmentation process, two colors are used to differentiate between lines' segments. Therefore, two aspects of measurements are involved; speed and accuracy.

The proposed Arabic OCR have been implemented in windows environment with visual C++ language. The first test has been done on laptop with Intel i7-3667U CPU, 2.0 GHz with 8 GB of memory and 256 GB hard drive. Table 6 listed the computational time using all the categories of the proposed dataset. The proposed Arabic OCR system is tested according to the universal Meta data description of the dataset for training. Comparing the speed of different processes (binarization, skewing and de-skewing, frame removing, noise removing, and segmentation) of the Arabic documents for different categories.

TABLE 6. SPEED PROCESSING (SECONDS)

IMAGE TYPE	PROCESSES SPEED (SECONDS)				
	BINARIZATION	SKEWING	NOISE REMOVING	FRAME REMOVING	SEGMENTATION
EARLY PRINTED	3.3	1.3	0.06	0.3	7
NEW BOOKS	2.7	0.7	0.04	0.3	5
CALLIGRAPHY	3.5	1.4	0.07	0.4	7.5
AVERAGE	3.16	1.11	0.056	0.33	6.5

6. Conclusion

In this paper, we presented datasets and metrics to be used in fair evaluation for existing methods in OCR or parts of systems. However, the need of public tools with common metrics and public datasets for evaluation arises according to the main characteristics of each dataset, and the dataset domains (types of images, fonts, styles, noise, etc.). Metrics for different modules in the OCR are discussed in this paper. Finally, ground-truth dataset for Arabic OCR, to be used as benchmarking, is presented. Creating, generating and then implementing standard evaluation tool to be used in OCR evaluation system have become strategic verification in the last year. Datasets creation is the starting point to go through this tunnel of imaged-documents analysis. Four Arabic datasets have been created in the different categories of the OCR tasks. In addition, various techniques for evaluation metrics are presented to be used in the proposed Arabic OCR system held in FCIT at KAU University. The most important metric was F-Measure, according to its usability at the three phases of the system. A new metric in the Arabic OCR system has been introduced for the offline recognition of different categories of cursive documents. Various technical methods in pre-processing for data preparation of the Arabic OCR system have been discussed, including PSNR, F-Measure, and negative rate metric (NRM). Experimental results showed that the proposed system achieves a good performance with 94.98 (for proposed system), 76.01 (for ABBYY Fine Reader) and 83.50 (for Google OCR), taken into consideration segmentation and recognition phases.

The recent availability of high processing power by using Graphical Processors (GPUs) has facilitated the adoption of advanced Deep Neural Networks Models for OCR systems which paved the way to improve the accuracy significantly. Also, the system speed became an important factor, since with this new hardware the system performance managed to run in real time.

Acknowledgment

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant no. (611-872-1435). The authors, therefore, gratefully acknowledge the DSR technical and financial support.

References

[1] Valveny E., Datasets and Annotations for Document Analysis and Recognition, David Doermann Karl Tombre Editors; Handbook of Document Image Processing and Recognition, Springer, Vol. 2, pp. 983-1011.

[2] Margner V. and Abed H. E., Tools and Metrics for Document Analysis Systems Evaluation, David Doermann Karl Tombre; Handbook of Document Image Processing and Recognition, Springer, Vol. 2, pp. 1010-1036.

[3] Gatos B., Imaging Techniques in Document Analysis Processes, David Doermann Karl Tombre Editors; Handbook of Document Image Processing and Recognition, Springer, Vol. 2, pp. 1011-1036.

[4] Alghamdi M., Alkhazi I., and Teahan W., Arabic OCR Evaluation Tool, 2016 7th International Conference on Computer Science and Information Technology (CSIT), 2016 IEEE.

[5] Saber S., Ahmed A., Hadhoud M., Robust Metrics for Evaluating Arabic OCR Systems, IEEE IPAS'14: International Image Processing Applications and Systems Conference 2014.

[6] Zayene O., Hennebert J., Touj S., Ingold R. and Ben Amara N., (2015). A Dataset for Arabic Text Detection, Tracking and Recognition in News Videos- AcTiV, 13th International Conference on Document Analysis and Recognition (ICDAR), pp 996-1000.

[7] Al-Barhamtoshy H., Fattouh A., Jambi K., Essa F., Khemakhem M., and Al-Ghamdi A., Universal Metadata Repository for Document Analysis and Recognition , The 13th ACS/IEEE International Conference on Computer Systems and Applications, AICCSA 2016.

[8] Al-Barhamtoshy H., Jambi K., Rashwan M., Abdou S., Sameer S., and Ahmed H. (2017). An OCR System for Arabic Calligraphy Documents, the International Conference on Communication, Management and Technology, ICCMIT 2017, University of Warsaw, Warsaw, Poland, April, 3-5, 2017.

[9] Pratikakis I, Gatos B, Ntirogiannis K (2011) Document image binarization contest (DIBCO 2011). International conference on document analysis and recognition (ICDAR 2011), Beijing, pp 1506–1510.

[10] Zhang Z. and Wang W. (2013). A Novel Approach for Binarization of Overlay Text, IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 4259-4264.

[11] Al-Barhamtoshy H. , and Rashwan A. Arabic OCR Segmented-based System, Life Science Journal, Vol. 11, Issue 11, (ISSN: 1097-8135), http://www.lifesciencesite.com .

[12] Attia, M., Arabic Orthography vs. Arabic OCR, Multilingual Computing & Technology magazine, USA, Dec. 2004.

[13] Zitouni I. (Editor), (2014). Natural Language Processing of Semitic Languages, Springer. Chapter 10: Darwish K, Information Retrieval, pp. 299-334.

[14] Nashwan F., Rashwan M., Al-Barhamtoshy H., Abdou S., and Moussa A., A Holistic Technique for an Arabic OCR System, Submitted to J. Imaging , pages 1 – 11.

[15] Alamri H, Sadri J, Suen CY, Nobile N (2008). A novel comprehensive database for Arabic offline handwriting recognition. In: Proceedings of the 11th international conference on frontiers in handwriting recognition (ICFHR 2008), Montreal, pp 664–669.

[16] Gatos B, Ntirogiannis K., Pratikakis I (2009) ICDAR2009 document image binarization contest. 10th International conference on document analysis and recognition (ICDAR'09), Barcelona, pp 1375–1382.

[17] Abdelaziz I., Abdou S., and Al-Barhamtoshy H., A large vocabulary system for Arabic online handwriting recognition, Pattern Analysis & Applications, Springer, Nov. 2016, 19(4), pp 1129-1141.

[18] Hesham A, Abdou S., Badr A., Al-Barhamtoshy H., Arabic Document Layout Analysis, Pattern Analysis and Applications, 2017, PAAA-D-15-00373R4. http://link.springer.com/article/10.1007/s10044-017-0595-x.

[19] Pletschacher S, Antonacopoulos A (2010). The page analysis and ground-truth elements format framework. In: 20th International conference on pattern recognition (ICPR), Istanbul, 2010, pp 257–260.

[20] Wang K, Belongie S (2010) Word spotting in the wild. In: Proceedings of the 11th European conference on computer vision: part I (ECCV'10), Heraklion. Springer, Berlin/Heidelberg, pp 591–604.

[21] Bukhari S, Shafait F, Breuel TM (2012) The IUPR dataset of camera-captured document images. In: Proceedings of the 4th international conference on camera-based document analysis and recognition (CBDAR'11), Beijing. Springer, Berlin/Heidelberg, pp 164–171.

[22] Zheng Y, Li H, Doermann D (2004) Machine printed text and handwriting identification in noisy document images. IEEE Trans PAMI 26(3):pp. 337–353.