IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

A Survey on the Existing Arabic Optical Character Recognition and Future Trends

Lutfieh S. Alhomed [1] and Kamal M. Jambi [2]

Lecturer, Computer Science and Software Engineering, University of Hail, Hail, Saudi Arabia[1]
Professor, Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia[2]

Abstract: Optical Character Recognition (OCR) is a computer system designed to transform images of typewritten text (typically captured by a scanner) into machine-processed text. Arabic OCR has been developed and enhanced over decades leading to the presence of enormous number of approaches with robust results that approaching accuracy, in some cases, of approximately 99%. However, existing OCRs exhibit shortages, or at least only sub-sets of them, when they were implemented with new applications, such as low-resolution inputs and video-based inputs. Accordingly, there is a need to review the existing approaches that showed robust results, analyses its mechanism and list its advantages and disadvantages in-order to ease the adaptation and extension of these systems into the new applications in this field. This paper presents a literature review on the existing systems for Arabic OCR, draw a common mechanism out of them, list the differences, advantages and disadvantages that helps in adaptation or extension of these systems to fit the recent demands.

Keywords: Optical Character Recognition, OCR, Pre-processing Operations, Segmentation, Optical Font Recognition.

I. INTRODUCTION

Optical Character Recognition (OCR) transforms images of typewritten text (typically captured by a scanner) into machine- processed text, which is usually encoded using representing scheme. OCR is the machine replication of human reading, which has been subject of intensive research since the foundation of computation devices. Arabic language, similar to other natural languages, such as English, has been investigated for OCR possibilities and robust Arabic OCR have been developed accordingly. However, Arabic differs from English as it is written with connected characters that possess extra challenging for Arabic compares to English. One of earliest surveys [a] deals with the topic where it covers researches in the late 80's. Each character in Arabic has a different shape depending on it position with respect to the word (i.e. start, middle, end, isolated). Also, different characters have the same main shape but they are different in number of dots and their location (e.g. Baa, Taa, and Thaa) [1]. As Arabic character segmentation is a necessary step in Arabic OCR, cursive nature of Arabic script poses challenges to this step and to Arabic OCR as a whole. This is because incorrectly segmented characters will cause misclassifications of characters and in turn will lead to inaccurate results. Arabic character segmentation is a difficult research problem due to both, the cursive nature of Arabic writing in both printed and handwritten forms and the scarcity of Arabic databases and dictionaries. This paper presents literature review of OCR of Arabic language. The framework for the presented literature review is organized as given in Figure 1. First, isolated character-based recognition is reviewed, which forms the core of the more general and applicable OCR. Then, the common process operations that preceded OCR are discussed in brief. A comprehensive OCR for Arabic is represented next. Based on the similarities and differences, a common OCR approach is concluded. The recent trends are then discussed with recommended solutions based on the discussion formed in throughout this paper.

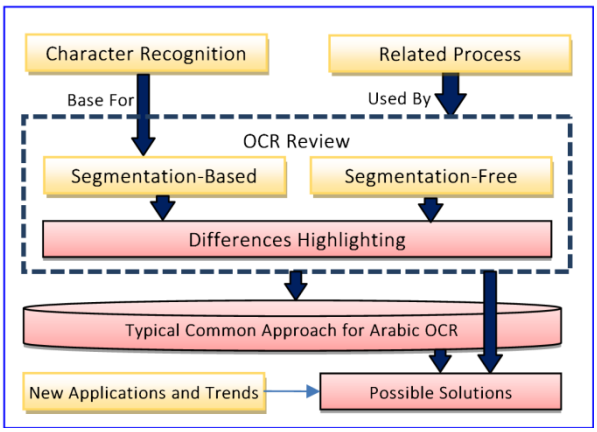

Figure 1: Literature Review and Recommendation Framework.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

II. CHARACTER RECOGNITION

Unlike most of existing Arabic OCR, some techniques have been developed to work with uncommon cases of isolated characters. Accordingly, character-based OCR is used to recognize a character in single character's image. For example, Kundu, et al. [2] extracted fifteen geometric and topological features, such as moments, zero-crossing and end-points, from the input image as input to Hidden Markova Model (HMM) [3] that is trained previously in-order to recognize the character in the image. Although such approach for character recognition is unlikely to be used as the goal of the OCR, which is developed to process real-life Arabic text that involves words and lines rather than isolated characters, this approach uses some common processing stages such as feature extraction and recognition using HMM that forms the core of OCR systems.

In fact, character recognition, as special case of OCR, have been studied extensively and various approaches have been developed in the literature. Generally these approaches uses features with a classifier, examples of these approaches in the literature, are: Dershowitz and Rosenberg [4], who used SIFT features [5] with HMM, Ali, et al. [6] who used principle component analysis [7] and neural network (NN) [8], Sahlol and Suen [9], who used structural, statistical and morphological features with NN and Fadel, et al. [10], who compare the results of various kernels with Support Vector Machine (SVM) [11].

III. PRE-PROCESSING OPERATIONS

In the process that preceded OCR, line segmentation and word segmentation are implemented on the input text image. Given that the expected input is normal Arabic text, it is necessary to implement word isolation using word and line segmentations, which are common processing units of most OCRs. Other pre-processing methods, such as normalization and skewness removal are also used sometimes with specific OCR systems.

III.1. LINE SEGMENTATION

Input text images for OCR is typically a printed page with multiple lines. To facilitate the utilization of OCR, line segmentation is implemented to isolate each text line for further processing. There are two approaches for line segmentation, these are: using fixed threshold as separation between the lines [12] or using horizontal projection [13]. The difficulty of line segmentation depends mainly on the type of the document (historical document, newspapers, books and letters) and the quality of the input document. Accordingly, various line segmentation methods have been proposed in the literature. Mohammed, et al. [14] computed the height of line segmentation and implemented line segmentation using line-height method that search for starting foreground pixel and end foreground pixel of each line. Ayesh, et al. [15] proposed a line segmentation approach that use the global maximum peak and baseline detection. This approach segments the input image into columns and each column is then segmented into rows, the line is then segmented by finding the maximum agreement between the rows of all the columns. Shakoori [16] proposed a line segmentation be delineating the foreground from the background, then the components of the foreground are connected using vertical and horizontal process. Noise is undesired disturbing variations of the image that affects the content of the image, accordingly, noise removal is required to ensure a correct processing and generated results. Various noise removal algorithms have been proposed and implemented for image noise removal, among these algorithms median filter [17] is often adapted in the existing OCR systems, however, some other filters, such as wiener2 filter can be used for Arabic OCR [18].

III.2. WORD SEGMENTATION

As lines are extracted from the input documents in the line segmentation process, each line is usually segmented into words using word segmentation process. Word segmentation faces two challenges, these are: the unconstraint layout and typesetting imperfections. Word segmentation is implemented by conducting virtually projecting the input into a histogram and determine the area with minimum foreground pixels [19]. Generally, word segmentation depends on determines the vertical white spaces between words which can be determined based on the pixels intensity. Gatos, et al. [20] proposed an efficient word spotting methodology using block-based document description. Similarly, various other methods used similar approach for word segmentation [21, 22]. Generally, image normalization is implemented to ease the process of feature extraction and content recognition. In OCR systems, regardless of the processed language, image normalization is implemented in form of binarization process by which the image of multiple colors or gray scales are converted into black and white images. The black is usually represented the text while the white represents the background. The binarization process is an easy step that implemented by determine a threshold value by which the image is binarized. Image thresholding is a well-studied field in image processing [23]. Besides normalization, other forms of image transformation are applied depends on the input data, such as thinning [24].

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

III.3. BASELINE DETECTION

As lines are extracted from the input documents in the line segmentation process, each line is usually segmented into words using word segmentation process. Word segmentation faces two challenges, these are: the unconstraint layout and typesetting imperfections. Word segmentation is implemented by conducting virtually projecting the input into a histogram and determine the area with minimum foreground pixels [19]. Generally, word segmentation depends on determines the vertical white spaces between words which can be determined based on the pixels intensity. Gatos, et al. [20] proposed an efficient word spotting methodology using block-based document description. Similarly, various other methods used similar approach for word segmentation [21, 22]. Generally, image normalization is implemented to ease the process of feature extraction and content recognition. In OCR systems, regardless of the processed language, image normalization is implemented in form of binarization process by which the image of multiple colors or gray scales are converted into black and white images. The black is usually represented the text while the white represents the background. The binarization process is an easy step that implemented by determine a threshold value by which the image is binarized. Image thresholding is a well-studied field in image processing [23]. Besides normalization, other forms of image transformation are applied depends on the input data, such as thinning [24].

III.4. SUMMARY

The processes, which are discussed in this section, differ marginally in the implemented techniques and the produced outputs. Example outputs of each of these processes, with an associated input are given in Figure 2.

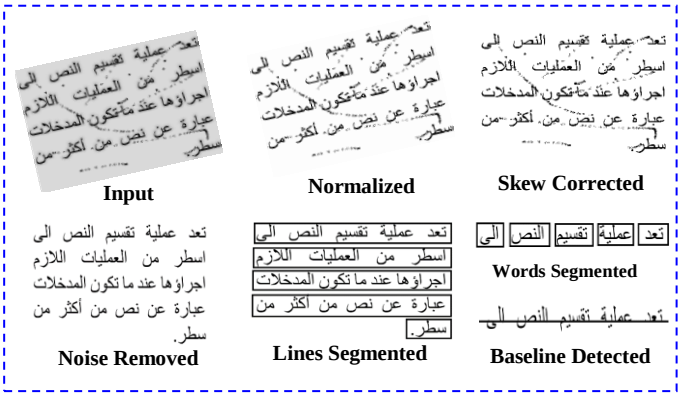

Figure 2: Examples Output of the Related Non-OCR Processes

As noted these processes are strongly related to the form of the input image, image quality, type (handwritten vs. typewritten) and length (full document, line, word and character) play a major role in determine to use or do not used each of these processes. For example, if the input is a line, line segmentation will not be used, while if the input is a word, neither line nor word segmentation will be used. Table 1 summarizes the cases of using each of these processes.

Table 1: Using Related Non-OCR Processes based on the Input Characteristics

Input	Line Seg.	Word Seg.	Baseline Det.	Noise Rem.	Norm.	Skew Corr.
Non-Skewed, Noise-Free, Binary Document	√	√	√	X	X	X
Non-Skewed, Noise-Free, Binary Line	X	√	√	X	X	X
Non-Skewed, Noise-Free, Binary Words	X	X	√	X	X	X
Non-Skewed, Noise-Free, Non-binary Document	√	√	√	X	√	X
Non-Skewed, Noise-Free, Non-binary Line	X	√	√	X	√	X
Non-Skewed, Noise-Free, Non-binary Words	X	X	√	X	√	X
Non-Skewed, Noisy, Binary Document	√	√	√	√	X	X
Non-Skewed, Noisy, Binary Line	X	√	√	√	X	X

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Non-Skewed, Noisy, Binary Words	X	X	√	√	X	X
Non-Skewed, Noisy, Non-binary Document	√	√	√	√	√	X
Non-Skewed, Noisy, Non-binary Line	X	√	√	√	√	X
Non-Skewed, Noisy, Non-binary Words	X	X	√	√	√	X
Skewed, Noise-Free, Binary Document	√	√	√	X	X	√
Skewed, Noise-Free, Binary Line	X	√	√	X	X	√
Skewed, Noise-Free, Binary Words	X	X	√	X	X	√
Skewed, Noise-Free, Non-binary Document	√	√	√	X	√	√
Skewed, Noise-Free, Non-binary Line	X	√	√	X	√	√
Skewed, Noise-Free, Non-binary Words	X	X	√	X	√	√
Skewed, Noisy, Binary Document	√	√	√	√	X	√
Skewed, Noisy, Binary Line	X	√	√	√	X	√
Skewed, Noisy, Binary Words	X	X	√	√	X	√
Skewed, Noisy, Non-binary Document	√	√	√	√	√	√
Skewed, Noisy, Non-binary Line	X	√	√	√	√	√
Skewed, Noisy, Non-binary Words	X	X	√	√	√	√

IV. OCR REVIEW

One of the earliest work deals with handwritten Arabic words is presented in [35]. The powerfulness of this technique comes from associating the structural features (such as end point, branch points, etc.) with the window number. This window is generated by the smallest rectangle that surround the Arabic isolated handwritten word. The shape of the window is modified in to a circle in [36].

Character connectivity style and the cursive nature of Arabic text are the main challenges of the Arabic OCR systems. Arabic characters are connected to each other with a reference that called “the baseline” to form words and sub-words. Some characters might connect to each other on the baseline and some are connected above the baseline as in ligatures. Generally, there are four different cases of characters connectivity with regarding to the baseline, these are: 1) No touching, 2) Ascender touching: the characters are connected above the baseline, 3) Descender touching: characters connected below baseline and 4) Baseline touching [37]. In order to deal with this variability in character connectivity, various approaches have been proposed. Overall, OCR systems is divided into two groups, these are: segmentation-based OCR and segmentation-free OCR.

IV.A.1 SEGMENTATION-BASED AOCR

The first approach deals with character connectivity problem by segmenting the characters in the word. Segmenting Arabic words into letters is the most difficult problem in Arabic script recognition [38]. Researchers have paid special attention to this problem and developed many algorithms to segment Arabic text into characters. Accordingly, various OCR systems have been proposed to address the segmentation challenges. As an example, Jambi [39] discusses the process of implementing an off-line system for segmenting and recognizing handwritten Arabic words. In order to recognize a word, its character decomposition should be known. Authors of [40] present an approach that makes use of geometric features extracted during the concurrent tracing of the upper and lower contours of the Arabic words. The selected features are precisely chosen to uniquely identify different character shapes. It should be mentioned that these features are not sensitive to scaling and thus the recognition results are more accurate.

El-Sheikh and El-Taweel [41] proposed a segmentation approach for OCR system based on dividing the Arabic characters into 4 groups of letters (initial, meddle, final and isolated) and implement segmentation based on the stroke. This approach would be extremely sensitive to noisy data in terms of number of strokes since the recognition system was built on counting the exact number of stroke. Similarly, Al-Emami and Usher [42] implemented a segmentation approach by dividing words into strokes by finding the extreme curvature, which was sensitive to rotation.

Elgammal, et al. [37] proposed an OCR system that used a graph-based segmentation to produce sub-character and a classifier to recognize these subs-character. The experiments were conducted on printed-text dataset and produced an average classification rate of 94.1%. The presented system was robust with characters connecting with baseline touching, while the proposed segmentation process failed in the other cases.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Cheung, et al. [43] proposed an OCR system that used a recognition-based segmentation technique to overcome the limitations in the classical segmentation approaches. A newly developed Arabic word segmentation algorithm is also introduced to separate horizontally overlapping Arabic words/sub-words. There is also a feedback loop to control the combination of character fragments for recognition purpose. The flowchart of this system is illustrated in Figure 3. The system was implemented and the results show a 90% recognition accuracy with a 20 chars/s recognition rate [43].

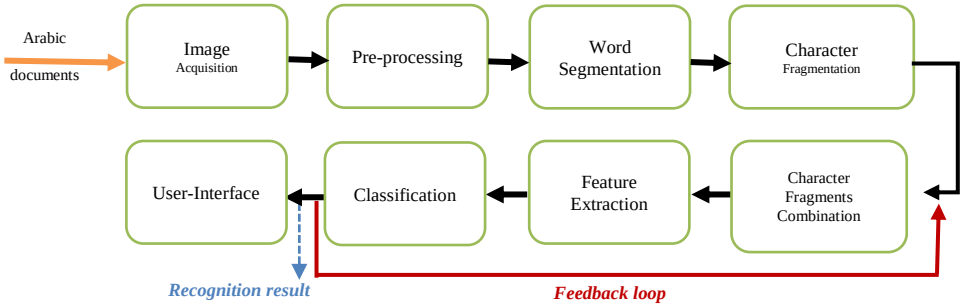

Figure 3: The Structure of the Proposed OCR System by Cheung, et al. [43]

Zheng, et al. [44] presented an OCR system that deal with the segmentation process in three-levels, these are: segment text into lines, segment each line into words, and segment each word into characters. Character segmentation is implemented based on the stroke of the characters. The segmentation algorithm has been tested and resulted in high accuracy rate of approximately 94%. Compared with the existing systems, this multi-level segmentation mechanism is very simple and efficient.

Schambach, et al. [38] proposed a segmentation approach for OCR system based on the facts that the connections between characters appear after an intersection or cusp points, or a change of curvature. But segmentation has to be proceed with successive essays and the character recognition module is responsible for validating each essay.

Daifallah, et al. [45] proposed a new OCR system that used over-segmentation of characters depending on the stroke. The segmentation process gives all possible segments that followed by segmentation enhancement, consecutive joints connection and finally segmentation point locating. The proposed system gives an excellent recognition rate up to 97% and 92% for words and letter recognition.

Razzak, et al. [46] presented an Arabic OCR system for Urdu script that is written in Nasta'liq and Naskh styles based on a novel technique for segmentation and recognition. First, the secondary strokes are segmented form the raw input strokes. Then, the primary baseline is extracted using the horizontal projection on ghost shapes. Finally, the local baseline of each ligature is estimated based on the extracted features and the estimated primary baseline. This approach obtained good results due to the utilization of baseline estimation over global reduction of diacritical marks. The proposed approach obtained an accuracy of 80.3% and 91.7% for Nasta'liq and Naskh font, respectively.

As has been noted, segmentation's challenges have been addressed by one of the following common directions: Pre-processing using base-line estimation, font resizing and etc., integrating sub-character segmentation with recognition, extreme-based segmentation.

IV.A.2 SEGMENTATION-FREE AOCR

Most of the existing Arabic OCR is of segmentation-based OCR, which is a successful approach in Latin typewriting, however, due to the obvious difficulties and errors resulted from the character segmentation process in Arabic language, some techniques have been proposed without segmentation-phase.

Bunke and Caelli [47] proposed an OCR system that decomposed the problem into a combination of two 1-D pattern recognition tasks. The goal of the first task, called line finding, is to locate the individual lines of text on a page. The goal of the second task, also called the recognition task, is to extract a set of features for the line and then use these features and the glyph HMMs to generate a sequence of characters or words for the line. A word or character language model (LM) is used to constrain the search. An advantage of the HMM based approach is that a line of text is not required to be pre-segmented into characters before recognition. The segmentation of the line into character boundaries is a by-product of the recognition process. For connected scripts, such as Arabic and cursive handwritten text, character segmentation is non-trivial and often in-accurate in the presence of noise artefacts. The segmentation free nature of HMM based OCR has resulted in widespread adoption of this approach [48, 49].

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Alma'adeed, et al. [50], used HMM for recognizing printed Arabic words of one hundred different writers. First, normalization processes were implemented without affecting the identity of the writer. Next, skeleton and edge of the word are determined and are used as features for the implemented system. Then, a classification process based on the HMM approach is used. Finally, the word extracted is compared with entities in the dictionary. This system was tested on a database of handwritten Arabic words and obtained an accuracy rate of about 45% and compared to other existing systems [51, 52].

Khorsheed [53] proposed a segmentation-free OCR system that is based on Hidden Markov Model (HMM) Toolkit (HTK). The proposed system consisted of pre-processing step using median filters that removes salt-and-paper noise and skew detection and skew correction. Besides, line segmentation is also performed. In the main processing stages, feature extraction from each line using overlapping vertical transformed window based on Sobel edge detection, is performed. Then, these features are passed into HMM with character model for character recognition. This system was tested on a dataset consists of Arabic text of more than 600 A4-size sheets typewritten in multiple computer-generated fonts. The effects of various parameters have been studied using this corpus and the system showed high capabilities in learning complex ligatures and overlaps. The system performance has been improved when implemented the tri-model scheme. The system performance showed an improvement from accuracy rate of 61.2%, when implementing mono-models, to accuracy rate of 85.9%, when implementing tri-models.

Prasad, et al. [54] proposed an HMM-based OCR system for machine-printed Arabic documents. A combination of script independent and script-specific techniques is applied to glyph models and language models (LM). The system was tested on machine-printed Arabic documents and showed relative reduction error rate of 40% on the baseline configuration.

Sabbour and Shafait [55] presented an OCR system for Arabic script languages called Nabocr. Nabocr were trained to recognize both Urdu Nastaleeq and Arabic Naskh fonts, however, it can be trained to be used for other Arabic script languages. In order to evaluate Urdu recognition, they have generated a dataset of Urdu text called UPTI (Urdu Printed Text Image Database), which measures different aspects recognition. The performance of their system for Urdu clean text was of rate up to 91%. For Arabic clean text, the performance was of rate up to 86%. Moreover, in comparison with, Tesseract [56], the performance of both systems on clean images is almost the same.

Al-Barhamtoshy and Rashwan [57] proposed an OCR system with various technical methods, in pre-processing, font resizing, binarization, de-skewing, denoising, and segmentation, are used. Experimental results showed that the proposed system achieves a good performance with less complexity. This system is illustrated in Figure 4.

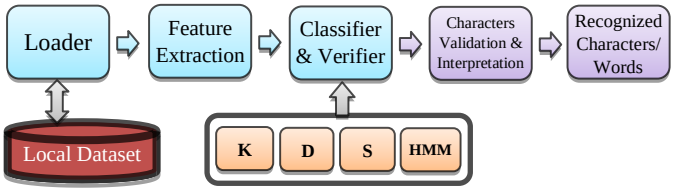

Figure 4: The Structure of the Proposed OCR System by Al-Barhamtoshy and Rashwan [57].

Khorsheed [58] proposed a segmentation-free OCR system using HTK with run-length encoding (RLE). RLE feature vectors ease the process of accurately fine-tune the recognition engine parameters and hence improve the overall system performance compared to the intensity feature vectors. In pre-processing stage, image resizing to 60 pixels height were implemented, which allow for consistent feature vector extraction. Various techniques were used to produce uniform feature vectors. The learning process focuses on complex ligatures and overlaps among words and characters. The performance of the proposed approach was tested using a corpus including cursive 600 A4-size pages Arabic typewritten text in six fonts, these are Tahoma, Simplified Arabic, Traditional Arabic, Andalus, Naskh, and Thuluth. Compared to existing intensity feature-based systems, this system produced significantly more accurate results.

The main drawback of segmentation-free approaches is the complexity especially for large vocabulary tasks. In this domain, Nashwan, et al. [59], proposed a computationally efficient holistic Arabic OCR. A lexicon reduction technique based on clustering similar shape words is utilized to reduce the word recognition time. The presented system used a combination of holistic global word-level and local block-based features. This system was tested using a dataset of 1152 words in three different fonts and four different sizes and has achieved an accuracy rate of 84%.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

No segmentation is adopted in the work of [60]. In this work, all structural features are detected for the whole word. Then scanning the image from right to left, and once the features of an Arabic character are detected, an Arabic character is indemnified where it is removed from the image as well as its associated features. This is followed by recognizing the next character. The process is repeated until all features are consumed which means that all characters of the Arabic word are recognized.

V. TYPICAL OCR FOR ARABIC LANGUAGE

Based on the discussed literature on Arabic OCR system, it can be concluded that the existing approaches share much in common. The six pre-processing steps discussed earlier are required for Arabic document process, accordingly, it has been used with most of the OCR systems. In post processing, language models, dictionaries and character variations are used to enhance the produced results. In the main processing steps, classification algorithms are the main component to be used. Figure 5 illustrates a general OCR system.

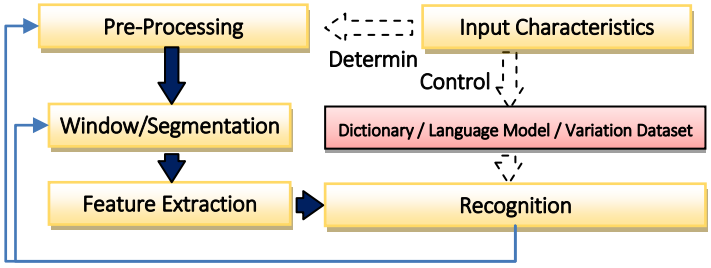

Figure 5: The Structure of the Common OCR Process

Overall, the common characteristics of the Arabic OCR systems can be summarized in the following points:

- Common pre-processing steps are required for every system in order to ensure accurate results.
- Variations of the pre-processing steps are tested based on different realization of the input documents.
- Segmentation-based approaches have lower accuracy and less capabilities compared to segmentation-free systems.
- Language models, dictionaries and character variations are required in order to ensure accurate results.

VI. RECENT APPLICATIONS AND TRENDS FOR ARABIC OCR

With the advances in OCR systems for Arabic language, various attempts have been made to use these OCR systems in more challenging applications, such as video-captured text and historical document processing. Similarly, with the realization of the significant of the font and size of the text on Arabic OCR, various approaches have been proposed to address this issue. These recent trends in Arabic OCR systems is discussed in this section.

VI.A.1 OPTICAL FONT RECOGNITION

Due to the variations among different Arabic font types, Arabic character recognition is still a challenge. Most literature consider only one font per text that results in low recognition accuracy when applied to another font. Accordingly, it is preferable to implement optical font recognition (OFR) priori to OCR followed by font-dependent OCR. Accordingly, Dahi, et al. [61] proposed an OCR system that used Optical Font Recognition (OFR) stage as a main process in the OCR system to automate the process of assigning each text font to a specific classifier tree. A comparative study of four recent algorithms of primitive AOCR has been performed to choose the best technique for each font [62]. Accordingly, a combining of statistical features has been proposed to train the, Random Forest Tree classifier. The proposed system was tested on the generated noise -free dataset (PAC-NF) and achieved a character recognition rate of 99.8-100%.

VI.A.2 FONT AND SIZE RECOGNITION

Besides font, the size of the written text also influenced the accuracy of Arabic OCR system. Accordingly, font and size recognition are implemented priori to OCR followed by font and size-dependent OCR, or most commonly, resizing text and use OFR followed by font-dependent OCR.

Slimane, et al. [63] proposed a simple and robust approach for font and size recognition in ultra-low-resolution Arabic text images. The proposed approach used Gaussian Mixture Models (GMMs) for estimating font, size, or font and size model likelihoods with respect to local features. Feature extraction is implemented using a fixed-length sliding window

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

from right to left. The main advantage of this approach is being segmentation-free. The proposed approach tested on APTI database and the results showed that character and word recognition error could be reduced by over 70% when OFR is implemented first, followed by font-specific OCR. Figure 6 illustrate the proposed approach by Slimane, et al. [63].

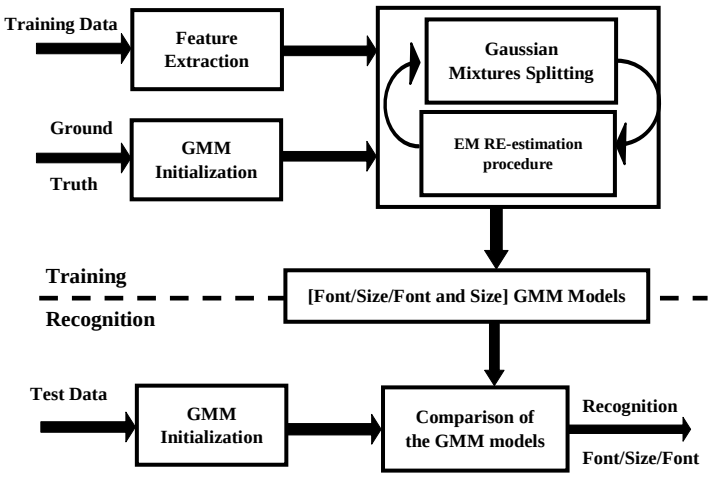

Figure 6: The Structure of the Proposed OCR System by Slimane, et al. [63]

VI.A.3 VIDEO-BASED OCR

Recent systems for text recognition in Arabic news captions have been proposed. The significant of such systems is to ease the process of video indexing and retrieval given that broadcast programs daily received and stored in big databases.

Yousfi, et al. [64] proposed a recognition approach that process input text image without any pre-processing operations or prior segmentation and implemented three methods for Arabic text recognition in the underlying videos. Each approach uses a deep learned model to represent the text image as a sequence of learned features. The experimental results showed that the proposed approach achieves good recognition rates that outperform the existing well-known commercial OCR systems.

Iwata, et al. [65] proposed an approach to detect and delineate the noisy moving news caption using OCR system based on inter-frame text difference to detect transition frame of still news captions. News caption recognition system consisted of text line extraction, word extraction and segmentation-recognition of words, which was evaluated using datasets of images extracted from AlJazeera broadcasting programs.

The technique was experimentally tested and shown to be robust to quick motion of the background and was able to detect the transition frame correctly with the F-measure higher than 90%. When compared with ABBY FineReader 11, commercial OCR the proposed OCR improves the recall of the Arabic character recognition from 70.74% to 95.85% for non-interlaced moving news caption images and from 23.82% to 96.29% for interlaced moving news caption images.

VI.A.4 HISTORICAL DOCUMENT OCR

Effective historical document indexing and retrieval poses a great challenge due to the vast amount of information that is available in libraries all over the world in the form of printed or handwritten manuscripts. The challenge is amplified by the variability of documents due to the multi-lingual and the wide range of historical periods that available collections are built, as well as by the poor quality of existing historical documents. OCR is required in order to facilitate advance indexing and retrieval of the contained information.

Adeyanju, et al. [66], proposed an OCR system that used HMM to recognize and classify degraded historical typewritten documents with broken edges, touching characters, broken characters, shape variance, skewing and heavy printing. Thus, this research proposed an efficient character recognition system using HMM as classifier. The results showed that this system has great effects of enhancing the accuracy of character recognition system. The result shows that recognition accuracy and precision for old memo dataset are 94.88% and 95%, for old war letter are 91.45% and

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

93%, for newly acquired dataset are 97.24% and 98% respectively, while FPR for old memo dataset is 0.0468 compared with 0.0736 for old war letter and 0.0206 for new typewritten essay dataset.

VII. RECOMMENDATION FOR RECENT TRENDS

The recent applications on OCR for Arabic language have poses new challenges for researchers in this field. However, based on the advances on OCR and the recent literature on these application, there are various possibilities to address these challenges, which are discussed in the following:

- Size Issues: The size problem of the input documents for AOCR can be directly solved by training the OCR on specific and medium size. Then, for recognition purpose, the recognition steps have to be preceded by size normalization, which can be increasing or decreasing using existing image processing techniques for image resizing.

- Font Issues: The font is different from the size. For a multi-font problem, a model has to be trained for each font. Then, for recognition purpose, the recognition steps have to be preceded by font detection, which can be implemented using commonly utilized classification algorithms, such as…? that used various features.

- Video Issues: Video captions problem for AOCR could be directly solved by correctly extracting the text from the video frames, which can be implemented using image segmentation approaches.

- Historical, Low quality and Low-Resolution Documents: Historical documents problem for AOCR is still an open problem. Transforming them to text is not easy problem. However, it could be handled by a process of annotation and associating these old documents with some labels that can be identifies automatically. On this type of document, they could be solved by means of using advanced image processing techniques to accurately extract some discriminating features from the inputs.

As such, most of the recent trends can use existing OCR systems, which showed good results with common inputs. However, before using these systems, the input documents, videos, historical documents have to be processed using some common image processing techniques for resizing, classification and image enhancement.

VIII. CONCLUSION

The Various aspects of the current OCR systems for Arabic language have been discussed in this paper. As noted, there is variations and harmony in the existing systems in various aspects. Mainly, there are six pre-processing steps that are (reorder according to the change in paper) noise removal, normalization, skew correction, baseline detection line segmentation and word segmentation, which are used regardless of the developed system but depend on the input characteristics. Similarly, some system-dependent steps are used as post processing operations to enhance the produced results. Those steps depend on processing the intermediate output. In the main processing steps, classification algorithms are the main component to be used. Nevertheless, the most obvious differences between the existing systems are whether it is segmentation-based or segmentation-free. Inter-process variations are embodied in the utilized character segmentation, windowing, and feature extraction and classification algorithms.

Recent applications of multi-font, multi-size, video captions and historical documents have poses new challenges and showed good possibilities. This is because most of the recent trends can used existing OCR systems with some common image processing techniques for resizing, classification and image enhancement.

REFERENCES

[1] K. Jambi, "ARABIC CHARACTER-RECOGNITION-MANY APPROACHES AND ONE DECADE," *Arabian Journal for Science and Engineering*, vol. 16, no. 4, pp. 499-509, 1991.
[2] A. Kundu, Y. He, and P. Bahl, "Recognition of handwritten word: first and second order hidden Markov model based approach," *Pattern recognition*, vol. 22, no. 3, pp. 283-297, 1989.
[3] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4-16, 1986.
[4] N. Dershowitz and A. Rosenberg, "Arabic Character Recognition," in *Language, Culture, Computation. Computing-Theory and Technology*: Springer, 2014, pp. 584-602.
[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, 1999, vol. 2, pp. 1150-1157: Ieee.
[6] A. Ali, A. Shaout, and M. Elhafiz, "Two stage classifier for Arabic Handwritten Character Recognition," *International Journal of Advanced Research in Computer and Communication Engineering*, pp. 646-650, 2015.
[7] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37-52, 1987.

[8] R. Hecht-Nielsen, "Theory of the backpropagation neural network," *Neural Networks*, vol. 1, no. Supplement-1, pp. 445-448, 1988.
[9] A. Sahlol and C. Suen, "A Novel Method for the Recognition of Isolated Handwritten Arabic Characters," *arXiv preprint arXiv:1402.6650*, 2014.
[10] S. Fadel, S. Ghoniemy, M. Abdallah, H. A. Sorra, A. Ashour, and A. Ansary, "Investigating the effect of different kernel functions on the performance of SVM for recognizing Arabic characters," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, pp. 446-450, 2016.
[11] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.
[12] T. Sari, L. Souici, and M. Sellami, "Off-line handwritten Arabic character segmentation algorithm: ACSA," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, 2002, pp. 452-457: IEEE.
[13] M. S. Khorsheed, "Off-line Arabic character recognition–a review," *Pattern analysis & applications*, vol. 5, no. 1, pp. 31-45, 2002.
[14] M. A. Mohammed, M. R. Kumar, and R. Pradeep, "Text Line Segmentation of Arabic Handwritten Documents using Line Height Method," *International Journal*, vol. 4, no. 11, 2014.
[15] M. Ayesh, K. Mohammad, A. Qaroush, S. Agaian, and M. Washha, "A Robust Line Segmentation Algorithm for Arabic Printed Text with Diacritics," *Electronic Imaging*, vol. 2017, no. 13, pp. 42-47, 2017.
[16] R. Shakoori, "A method for text-line segmentation for unconstrained Arabic and Persian handwritten text image," in *Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on*, 2014, pp. 338-344: IEEE.
[17] Y. Zhu and C. Huang, "An improved median filtering algorithm for image noise reduction," *Physics Procedia*, vol. 25, pp. 609-616, 2012.
[18] A. Sharma and D. R. Chaudhary, "Character recognition using neural network," *International Journal of Engineering Trends and Technology (IJETT)-Volume4*, pp. 662-667, 2013.
[19] K. Anwar and H. Nugroho, "A segmentation scheme of arabic words with harakat," in *Communication, Networks and Satellite (COMNESTAT), 2015 IEEE International Conference on*, 2015, pp. 111-114: IEEE.
[20] B. Gatos, I. Pratikakis, A. th International Conference on Document, and Recognition, "Segmentation-free Word Spotting in Historical Printed Documents," (in No Linguistic Content), pp. 271-275, 2009.
[21] H. Al-Rashaideh, "Preprocessing phase for Arabic word handwritten recognition," *Information Process (Russian)*, vol. 6, no. 1, 2006.
[22] J. H. AlKhateeb, J. Jiang, J. Ren, and S. Ipson, "Component-based segmentation of words from handwritten Arabic text," *International Journal of Computer Systems Science and Engineering*, vol. 5, no. 1, 2009.
[23] A. A. A. Shareha, M. Rajeswari, and D. Ramachandram, "Textured Renyi Entropy for Image Thresholding," in *Computer Graphics, Imaging and Visualisation, 2008. CGIV'08. Fifth International Conference on*, 2008, pp. 185-192: IEEE.
[24] S.-C. Pei and C.-N. Lin, "Image normalization for pattern recognition," *Image and Vision computing*, vol. 13, no. 10, pp. 711-723, 1995.
[25] B. Parhami and M. Taraghi, "Automatic recognition of printed Farsi texts," *Pattern Recognition*, vol. 14, no. 1-6, pp. 395-403, 1981.
[26] B. Timsari and H. Fahimi, "Morphological approach to character recognition in machine-printed Persian words," in *Document Recognition III*, 1996, vol. 2660, pp. 184-192: International Society for Optics and Photonics.
[27] M. Pechwitz and V. Margner, "Baseline estimation for Arabic handwritten words," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, 2002, pp. 479-484: IEEE.
[28] F. Farooq, V. Govindaraju, and M. Perrone, "Pre-processing methods for handwritten Arabic documents," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, 2005, pp. 267-271: IEEE.
[29] A.-K. Arwa, S. A. Pitchay, and M. Al-qudah, "An Arabic Baseline Estimation Method Based on Feature Points Extraction," in *Proceedings of the World Congress on Engineering*, 2017, vol. 1.
[30] H. Boukerma and N. Farah, "A novel Arabic baseline estimation algorithm based on sub-words treatment," in *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, 2010, pp. 335-338: IEEE.
[31] I. S. Abuhaiba, "Skew correction of textural documents," *Journal of King Saud University-Computer and Information Sciences*, vol. 15, pp. 73-93, 2003.
[32] M. Sarfraz and Z. Rasheed, "Skew estimation and correction of text using bounding box," in *Computer Graphics, Imaging and Visualisation, 2008. CGIV'08. Fifth International Conference on*, 2008, pp. 259-264: IEEE.
[33] A. Boukharouba, "A new algorithm for skew correction and baseline detection based on the randomized Hough Transform," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 1, pp. 29-38, 2017.
[34] R. N. Verma and L. G. Malik, "REVIEW ON SKEW DETECTION AND CORRECTION TECHNIQUES."
[35] K. Jambi, "Design and implementation of a system for recognizing Arabic handwritten words with learning ability," 1991.
[36] K. M. JAMBI, "An Experimental approach for recognizing handwritten Arabic words," *Science*, vol. 7, no. 1, 1995.
[37] A. M. Elgammal, M. A. Ismail, A. Proceedings of Sixth International Conference on Document, and Recognition, "A graph-based segmentation and feature extraction framework for Arabic text recognition," (in No Linguistic Content), pp. 622-626, 2001.
[38] M.-P. Schambach, J. Rottland, and T. Alary, "How to convert a Latin handwriting recognition system to Arabic," *ICFHR*, vol. 8, pp. 265-270, 2008.
[39] K. M. Jambi, "An approach for segmenting handwritten arabic words," in *Langue Arabe et Technologies Informatiques Avancées. Actes du colloque organisé par la Fondation du Roi Abdulaziz Al Saoud pour les Etudes Islamiques et les Sciences Humaines. Casablanca*, 1993, vol. 8, pp. 233-243.
[40] "<G. Abo-Samara and K. Jambi, Using Contour-Based Features for Segmenting and Recognizing Handwritten Arabic Words in Off-Line Systems, Journal of Engineering and Applied Science, Vol. 46, No. 2, April 1999..PDF>."
[41] T. S. El-Sheikh and S. G. El-Taweel, "Real-time arabic handwritten character recognition," (in English), *PR</cja:jid> Pattern Recognition*, vol. 23, no. 12, pp. 1323-1332, 1990.
[42] S. Al-Emami and M. Usher, "On-line recognition of handwritten Arabic characters," (in No Linguistic Content), *IEEE Trans. Pattern Anal. Machine Intell. IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 704-710, 1990.
[43] A. Cheung, M. Bennamoun, and N. W. Bergmann, "An Arabic optical character recognition system using recognition-based segmentation," (in English), *PR</cja:jid> Pattern Recognition*, vol. 34, no. 2, pp. 215-233, 2001.
[44] L. Zheng, A. H. Hassin, and X. Tang, "A new algorithm for machine printed Arabic character segmentation," (in English), *Pattern Recognition Letters Pattern Recognition Letters*, vol. 25, no. 15, pp. 1723-1729, 2004.
[45] K. Daifallah, N. Zarka, H. Jamous, A. th International Conference on Document, and Recognition, "Recognition-Based Segmentation Algorithm for On-Line Arabic Handwriting," (in No Linguistic Content), pp. 886-890, 2009.
[46] M. I. Razzak, M. Sher, and S. A. Hussain, "Locally baseline detection for online Arabic script based languages character recognition," (in English), *Int. J. Phys. Sci. International Journal of Physical Sciences*, vol. 5, no. 7, pp. 955-959, 2010.
[47] H. Bunke and T. Caelli, *Hidden Markov models : applications in computer vision*. Singapore; River Edge, NJ: World Scientific, 2001.
[48] P. Natarajan, S. Saleem, R. Prasad, E. MacRostie, and K. Subramanian, "Multi-lingual Offline Handwriting Recognition Using Hidden Markov Models: A Script-Independent Approach," (in English), *Lecture notes in computer science.*, no. 4768, pp. 231-250, 2008.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

[49] H. Bunke, S. Bengio, and A. Vinciarelli, "Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models," (in English), *IEEE transactions on Pattern analysis and Machine intelligence*, vol. 26, no. 6, pp. 709-720, 2004.
[50] S. Alma'adeed, C. Higgens, D. Elliman, and R. Proceedings of 16th International Conference on Pattern, "Recognition of off-line handwritten Arabic words using hidden Markov model approach," (in No Linguistic Content), vol. 3, pp. 481-484 vol.3, 2002.
[51] A. Amin and J. F. Mari, "Machine recognition and correction of printed Arabic text," (in No Linguistic Content), *IEEE Trans. Syst., Man, Cybern. IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 5, pp. 1300-1306, 1989.
[52] M. Dehghan, K. Faez, M. Ahmadi, and M. Shridhar, "Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM," (in English), *Pattern recognition.*, vol. 34, pp. 1057-1066, 2001.
[53] M. S. Khorsheed, "Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK)," (in English), *PATREC Pattern Recognition Letters*, vol. 28, no. 12, pp. 1563-1571, 2007.
[54] R. Prasad, S. Saleem, M. Kamali, R. Meermeier, and P. Natarajan, "Improvements in Hidden Markov Model Based Arabic OCR," (in English), *Proceedings /*, vol. 2, no. Conf 19, pp. 769-772, 2008.
[55] N. Sabbour and F. Shafait, "A segmentation-free approach to Arabic and Urdu OCR," (in English), *Document Recognition and Retrieval*, vol. 8658, pp. 86580N-86580N-12, 2013.
[56] R. Smith, "An overview of the Tesseract OCR engine," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, 2007, vol. 2, pp. 629-633: IEEE.
[57] H. Al-Barhamtoshy and M. Rashwan, "The Arabic OCR Segmented-based System," *Life Science Journal*, vol. 11, no. 10, 2014.
[58] M. S. Khorsheed, "Recognizing cursive typewritten text using segmentation-free system," *The Scientific World Journal*, vol. 2015, 2015.
[59] F. Nashwan, M. A. Rashwan, H. M. Al-Barhamtoshy, S. M. Abdou, and A. M. Moussa, "A Holistic Technique for an Arabic OCR System," *Journal of Imaging*, vol. 4, no. 1, p. 6, 2017.
[60] "<Gibrael Al Amin Abo Samra and Kamal Jambi, "Using contour-based features for recognizing handwritten Arabic words in off-line systems", Journal of Engineering and Applied Science, Vol. 46, No. 2, April 1999..PDF>."
[61] M. Dahi, N. A. Semary, and M. M. Hadhoud, "Primitive Printed Arabic Optical Character Recognition using Statistical Features," in *Intelligent Computing and Information Systems (ICICIS), 2015 IEEE Seventh International Conference on*, 2015, pp. 567-571: IEEE.
[62] M. Rashad and N. A. Semary, "Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers," in *International Conference on Advanced Machine Learning Technologies and Applications*, 2014, pp. 11-17: Springer.
[63] F. Slimane, S. Kanoun, J. Hennebert, A. M. Alimi, and R. Ingold, "A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution," (in English), *Pattern Recognition Letters Pattern Recognition Letters*, vol. 34, no. 2, pp. 209-218, 2013.
[64] S. Yousfi, S.-A. Berrani, and C. Garcia, "Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, 2015, pp. 1026-1030: IEEE.
[65] S. Iwata, W. Ohyama, T. Wakabayashi, and F. Kimura, "Recognition and transition frame detection of Arabic news captions for video retrieval," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, 2016, pp. 4005-4010: IEEE.
[66] I. Adeyanju, O. Ojo, and E. Omidiora, "Recognition of typewritten characters using hidden Markov models," *British Journal of Mathematics & Computer Science*, vol. 12, no. 4, p. 1, 2016.