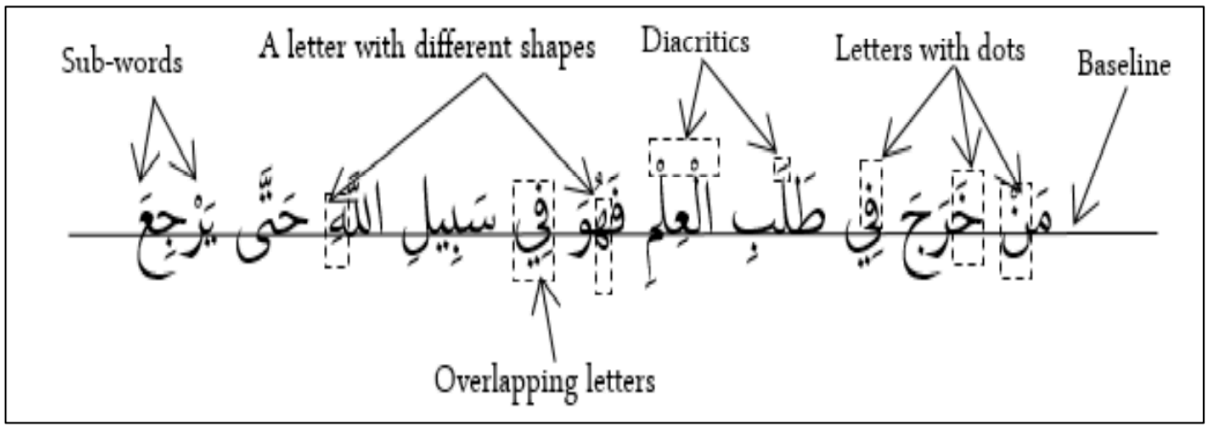
Printed Arabic Script Recognition

This paper has overviewed the main stages used in printed Arabic OCR. It main aim is to reveal the current status of printed Arabic OCR

ARABIC SCRIPT CHARACTERISTICS AND CHALLENGES



ARABIC CHARACTERS WITH DIFFERENT POSITIONS AND SHAPES

Isolated	Initial	Middle	End
ا	ا	ـا	ـا
ب	بـ	ـبـ	ـب
ت	تـ	ـتـ	ـت
ث	ثـ	ـثـ	ـث
ج	جـ	ـجـ	ـج
ح	حـ	ـحـ	ـح
خ	خـ	ـخـ	ـخ
د	د	ـد	ـد
ذ	ذ	ـذ	ـذ
ر	ر	ـر	ـر
ز	ز	ـز	ـز
س	سـ	ـسـ	ـس
ش	شـ	ـشـ	ـش
ص	صـ	ـصـ	ـص
ض	ضـ	ـضـ	ـض
ط	طـ	ـطـ	ـط
ظ	ظـ	ـظـ	ـظ
ع	عـ	ـعـ	ـع
غ	غـ	ـغـ	ـغ
ف	فـ	ـفـ	ـف
ق	قـ	ـقـ	ـق
ك	كـ	ـكـ	ـك
ل	لـ	ـلـ	ـل
م	مـ	ـمـ	ـم
ن	نـ	ـنـ	ـن
ه	هـ	ـهـ	ـه
و	و	ـو	ـو
ي	يـ	ـيـ	ـي

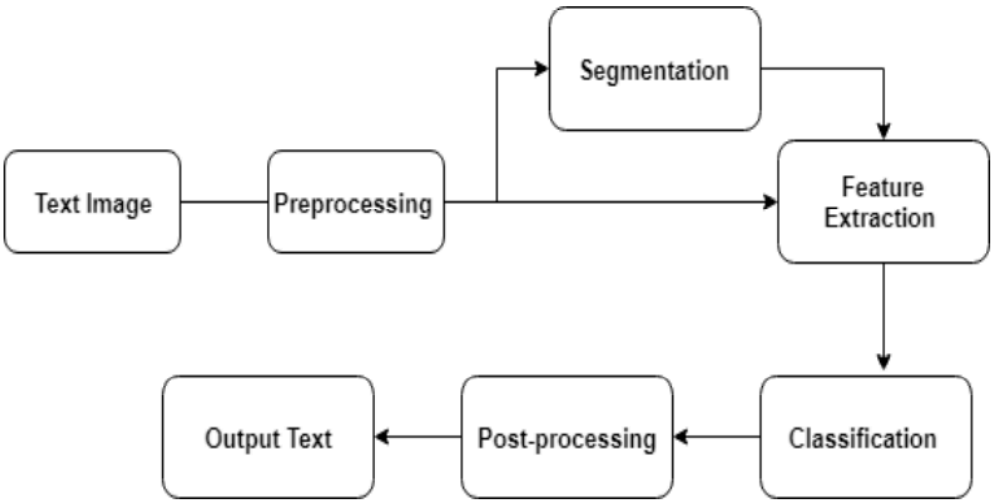
the characteristics of Arabic script that may complicate recognition will be discussed:

A. Shapes and Positions
B. Overlapping characters and Ligatures
C. Diacritics
D. Cursive
E. Presence of dots

GENERAL ARABIC OCR METHODOLOGY (MODEL):

Published approaches and systems for Arabic OCR indicate that the process of implementing Arabic OCR consists of five phases:

1. preprocessing.
2. segmentation
3. feature extraction
4. classification
5. post-processing



Preprocessing Phase

This is the first phase of OCR methodology which is responsible for enhancing the readability of the input image.

Preprocessing is a combination of algorithms that are applied to the input image in order to reduce noise and alterations, thus simplifying the subsequent phases of OCR methodology. There are various factors that affect the quality of the input image. A study lists the history of image, the printing process, the kind of font, the quality of paper, the condition of the image and the image acquisition as the vital factors that influence the input image quality.

Generally, several preprocessing operations are employed on the input image: binarization, layout analysis, thinning, smoothing and filtering, size and slant normalization, slant detection, skew detection and baseline detection.

However, the selection of these operations, to be applied in the preprocessing, relies upon the conditions of the input image, such as the amount of noise and skew in the input image.
the preprocessing techniques which are applied in Arabic OCR:

1. Binarization (thresholding)
   converting an input gray scale image into a binary image, in which a pixel has only two values 0 and 1
2. Size Normalization
   size normalization is commonly applied to characters or words by scaling the characters or the words to an adjusted size.
   This process is crucial for the recognition or classification phase
   A study classified normalization methods into two approaches:
   - moment-based normalization
   - nonlinear normalization
3. De-noising
   Noise may have a major impact on the performance of OCR systems.
   Noise removal is an operation for enhancing the visual quality of the input image.
   several techniques have been introduced that are considered as noise removal methods:
   - filtering
     the median filter approach is commonly used in both printed text images and handwritten text images
   - morphological operations (smoothing)
     - ★ dilation algorithms, which are applied to broken letters
     - ★ erosion algorithms which are applied to text images with touching letters
4. Skew Detection and Correction
   Initially, a text image has zero rotation, yet when physically scanning the image manually, rotation of images up to 20^{o} might occur. This rotation is called skew which results in non-zero skew text images.
   The skew can lead to incorrect recognition and baseline detection.
   It is impossible to segment a text if the text is rotated.
   The process of estimating the skew angle is known as skew detection
   the process of rotating the image with the purpose of correcting the skew is called skew correction
   A wide variety of skew detection and correction methods have been proposed.
   A study groups these methods into five groups:
   - projection profile

- Hough transform
- Fourier transform
- nearest neighbor clustering
- correlation

5. Baseline Detection
   Arabic characters are joined through a horizontal line called the baseline.
   Graphically, the baseline can be described as the line which has the maximal amount of black pixels.
   This line contains critical information about the text, such as text orientation and position of connection points between Arabic letters.
   The baseline detection techniques for Arabic script has been classified into four groups in :
   - namely
   - horizontal projection methods
   - the word skeleton method
   - contour tracing and principal component analysis

   baseline approches:
   1. horizontal projection approach
      widely implemented for determine the baseline in Arabic OCR
      The horizontal projection method is simple and efficient for Arabic printed text.
      However, this method is applicable only for noise-free images, as it fails for unclean images.
   2. the x-y cut approach
      This method works well for Arabic noisy images, though it fails in the presence of large amounts of noise and skew.

6. Thinning and Skeletonization
   the process of peeling off a pattern as many pixels as possible without affecting the general shape of the pattern
   In other words, it involves operations that can be implemented in order to produce the skeleton of text images.
   Thinning is a crucial processing step for text recognition.
   When applying thinning algorithms to Arabic scripts, various obstacles are encountered. One problem is the reduction in the number of dots in some Arabic characters as a result of the thinning process for which the number of dots is a crucial aspect in differentiating between these characters.
   Also, dots in Arabic characters are likely to be vulnerable to noise

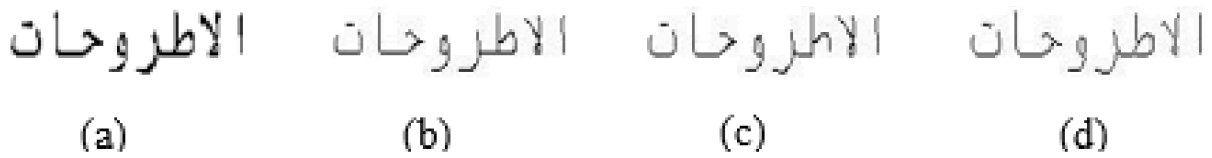


Fig. 6. Example results of different thinning algorithms: (a) original word, (b), (c) and (d) thinned word.

B. Segmentation Phase

During the segmentation phase, the text image is segmented into small components, with a page being segmented into lines, a line into words and a word into letters. segmenting a text image can be graded into two types:

- external segmentation
- internal segmentation

Fig. 8. Segmenting Arabic words into their characters.

External Segmentation

External segmentation refers to the document layout analysis, in particular page decomposition.

Generally, it is relatively easy to segment a text line into words in printed text images, compared to handwritten text images which involve overlapping and touching characters by using vertical projection histogram profiles

Internal segmentation

Internal segmentation deals with segmenting a word into characters. When reviewing segmentation methods in the literature, a major complication arises concerning the classification of word segmentation approaches.

Arabic OCR systems have been developed by two main paradigms:

- holistic approaches (segmentation–free) which require a large lexicon of Arabic words
- analytical approaches (segmentation based) where a word is segmented into units and each unit is recognized separately

Holistic Approach:

Segmentation-free or holistic Arabic OCR systems perform the recognition of the entire word as a unit without segmenting the word or recognizing characters separately.

Analytical Approach:
For the analytical or segmentation based approach, Arabic OCR systems segment words into smaller units like characters.
analytical approach is divided into two approaches: explicit segmentation and implicit segmentation.

- Explicit Segmentation (dissection segmentation)
  attempts to segment a word into smaller units. These units could be characters, strokes or loops.
  there are two classes of explicit segmentation, which are:
  - ★ direct segmentation
  - ★ indirect segmentation

- Implicit Segmentation
  In OCR systems based on implicit segmentation, the segmentation phase and recognition phase are performed simultaneously. In other words, a word is segmented into characters while being recognized without segmentation in advance . Straight segmentation and recognition based segmentation are also referred to as implicit segmentation.

C. Feature Extraction Phase
the process of obtaining distinguishing attributes of the segmented character to be utilized by the next phase which is classification.
The authors point out that the selection of feature extraction methods depends on the output of the preprocessing stage.
the set of features extracted must match the specification of the selected classifier.
However, selection of feature types is a major issue in OCR development.
Such features can be categorized into three groups:

- structural features
- statistical features
- global transformation feature

1. Structural Features
   Structural features illustrate a text image in terms of its topological and geometrical characteristics by using its local and global properties.

2. Statistical features
   Statistical features are derived from statistical representation of patterns which provide a measurable event of interested patterns.
   different approaches to produce statistical features. Some examples of the approaches are zoning, moments, characteristic loci, histograms and crossing.

3. Global transformation feature
   The global transformation method is applied to convert a skeleton or contour of a pattern by a linear transform into a form that reflects the most relevant features of the transformed pattern.

In conclusion, the feature extraction stage plays a critical role in Arabic OCR development in which distinguishing attributes are extracted and it is clear that each Arabic OCR developer needs to apply different feature extraction approaches. Still, good features are required, which assist in distinguishing a character from other characters and maximize the accuracy performance simultaneously. Furthermore, these features must be selected specifically for a selected classifier. Some researchers apply different feature extraction methods in combination. However, this may cause extra complications for the implementation.

D. Classification Phase

The classification phase has the responsibility for assigning a pattern into a pre-classified class based on the features of the pattern which have been extracted in the previous phase.
The pre-classified classes can be words, sub-words, characters or strokes, based on the OCR approach used. There are a number of different classification approaches that have been applied for Arabic OCR, such as Hidden Markov Models (HMM), Support Vector Machines (SVM), K-nearest neighbour(KNN).

E. Post-processing Phase

Post-processing is the final stage of the development of Arabic OCR
The objective of this step is to enhance the recognition accuracy by detecting and correcting linguistic misspellings in the produced OCR text without human intervention
Generally, post processing methods can be categorized into two main approaches:

- lexicon-based methods
- context-based (statistical) methods

The typical technique for correcting the mistakes of Arabic OCR outputs is the lexicon-based method which requires the utilization of an Arabic dictionary.
This technique corrects errors without considering any contextual information in which the errors appear

PERFORMANCE EVALUATION

OCR performance evaluation can be classified into two types:

- black-box evaluation
- white-box evaluation

Performance evaluation of OCR systems is essential for:

- monitoring progress of OCR systems development
- assessing the effectiveness of OCR algorithms
- identifying open areas for further research
- providing scientific justification for the performance of OCR systems

For Arabic OCR, conducting performance evaluation is challenging as no standard dataset is available.
This accuracy metric is insufficient to assess how Arabic OCR systems are overcoming the challenges of Arabic text. However, a study suggests a new set of objective performance metrics for evaluation Arabic OCR with respect to the challenges of Arabic script which are character accuracy based on character position, dot character accuracy, zigzag-shaped character accuracy, loop-shaped character accuracy and diacritics accuracy.