ID Card Recognition Based on Arabic OCR System

Amira Abdel-Kareem[*1], Ashraf Hussein[*2], Esraa Shokry[*3], OlaAlaa El-Din[*4,]

Mohsen. A. Rashwan[*5], and Hassanin M. Al-Barhamtoshy[**6]

[*] *Electronics and Communication Department, Faculty of Engineering, Cairo University*
Giza, Egypt
[1] amira_shaban22@yahoo.com
[2] h_ashraf16@hotmail.com
[3] engesraa_2013@hotmail.com
[4] olaalaa_2013@yahoo.com
[5] mrashwan@rdi-eg.com
[**] *Faculty of Computing & Information Technology, King Abdulaziz University*
Jeddah, Saudi Arabia
[6] hassanin@kau.edu.sa

Abstract- ***Optical character recognition of Arabic language is a field of research that is socially very relevant and challenging. The social relevance lies on the fact that OCR is very important for many applications that need character recognition of images. Our system is Egyptian ID cards reader system which extracts important data from the ID card image, recognizes data and translates it into editable text on computer so it can be edited and saved. Then, the system can compare between a tested ID card and the database saved before. This paper extensively reviews the base line-based segmentation and DCT based feature extractor approaches used for building this special Arabic OCR system. It also reports the experimental results obtained so far showing the reliability of our system. Finally, we'll show thatthe system works fast on the scientific Matlab library, as it needs about 16 seconds in average to process one ID card, and the system is expected to do better performance when transferring it from the academic phase to the product phase.***

Keywords: ***Arabic OCR, Card Identification, Information Retrieval, Classifiers Fusion***

1 INTRODUCTION

Humans recognize characters easily and they repeat the character recognition process thousands of times every day as they read papers or books [1]. Though, after many years of serious investigation and research, the ultimate goal of developing an optical character recognition (OCR) system, with the same interpretation capabilities as humans, still remains unachieved. One of the main objectives of an OCR is to reach a speed of 5 characters/second with a 99.9% recognition rate, with no errors.

The OCR is the electronic translation of scanned images of typewritten, printed text, or handwritten, into electronic mechanism encoded text [2]-[5]. OCR allows the machine automatically to recognize characters in an image and translate them into computer textual format by applying machine learning mechanism.

Therefore, development of the OCR systems is a very significant field of research in pattern recognition [9], [10]. Different OCR engines allow the machine to automatically recognize characters in an image and translate them into computer textual format by applying machine learning mechanism. This improves human-machine interaction and is widely used in many areas [6], [9]. Here's a general (OCR) DFD diagram, as shown in Fig. 1.

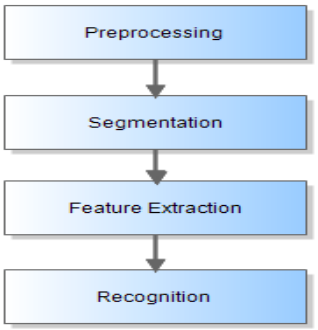


Figure 1: OCR DFD Diagram

Recently, identification cards recognition systems became very important due to the automatic processes of checking and storing card's data achieved by them in many applications, such as election systems and vital governmental installations, and making this operation easy and fast using optical solutions. So, from the realization of the importance of such

applications and their impact on the society, this paper shows developing an Arabic OCR system dedicated for ID card recognition.
The most common information that is used in Egyptian ID card includes: Name, Address, Birthdates, Job title, Gender, Religion, Marital status, Husband Name, Release date, and Expire date.
Arabic script characteristics and Arabic OCR challenges are discussed in section 2. An overview of the system will be presented in section 3. The proposed OCR algorithm will be discussed in section 4.Experimental work and results are offered in section 5. Finally, conclusion and future work will be discussed in section 6.

2 ARABIC SCRIPT CHARACTERISTICS AND OCR CHALLENGES

Arabic is the official language of over twenty Arab countries which stretch from Morocco to Iraq, it is the religious language of all Muslims, more than one billion Muslims spread all over the world, and it is the language of the Quran (the sacred book of Islam). Arabic language is a Semitic language and most of its words are built up from roots by following specific morphological grammatical rules by employing affixes processing (infixes, prefixes and suffixes). Classical Arabic language is widely used around fourteen centuries ago.
Arabic is a popular script and cursive nature language. More than one billion Arabic script users are estimated in the world. Due to the cursive nature of Arabic text, the development of Arabic OCR systems involves much technical difficulty, mainly in the segmentation phase. Although many researchers are studying solutions to solve such troubles, very little progress has been made [3]-[5].

A. Arabic Script Characteristics

Arabic is written from right to left, and so the recognition process should occur from right to left. Arabic character set includes 28 letters; each Arabic letter has 2-4 different forms which depend on its position in the word. Fifteen of the 28 letters have dots and the other 13 are without dots. Dots are above or below the Arabic letters, it plays a major role in discriminating some characters from each similar, that differ only by the number or location of dots; e.g. letters (ب- ت- ث- ن- ي). There are four characters which may coupled the letter "Hamzah ء", those are "Alif , أ إ ", "Waw ؤ" , "Yaa ئ" and "Kaf" ك. Six Arabic letters can be connected from the right side only: dal (د), raa (ر), waw(و), alef (ا), thal (ذ), and zay(ز). While the other 22 letters, can be connected from both sides. These six letters have just two forms, the stand-alone form and the final form [4], [5].

Arabic letters do not have fixed width or rigid size, even in printed documents. The shape of the letter is influenced by its position in the word. Whereas the rest of the characters can appear in any of four shapes: the beginning, the middle, the ending, and the separate form. Consequently, an Arabic word may consist of one or more sub-words. A sub-word can be defined as the basic separate pictorial block of the Arabic writing.

Any OCR system of Arabic characters should take care of the sub-word as the basic segment for processing. This is because each sub-word is separated from other sub-word by a gap. Although. spaces between sub-words are regularly shorter than those between consecutive words. A word may contain one or more sub-words, and some of these sub-words could even consist of a single character in its separate form. Accordingly, their recognition does not need segmentation. Shape of the letter in the text differs according to the location of the character in the sub-word, i.e. a character at the end of sub-word, has exactly the same shape when it comes at the end of a full word.

The Arabic character set is shown in Table I, that illustrates the variation of the Arabic characters' shape depending on their positions in the word [3], [4].

B. Arabic OCR Challenges

Although working on a standard ID has advantages like, the font type is fixed (simplified Arabic font) and approximately similar font sizes, Arabic words may horizontally overlap and could not be separated; i.e., letters may stack on others. These introduce troubles for both the word and the character segmentations. At this level, it is not hard to understand that segmentation is a crucial step in the development of an Arabic OCR system. The main difficulty associated cursive text recognition is the segmentation of words to letters. Following is an introduction to the main challenges in Arabic OCR system [7], [10].

1) Connectivity Challenge: As illustrated before that the cursive phenomenon of Arabic language text introduces problems for both the word and the character segmentations. The main difficulty associated with cursive text recognition is the segmentation. Accordingly, the segmentation is a critical step in the development of an Arabic OCR system. Fig. 2 illustrates Arabic naming script showing connectivity.

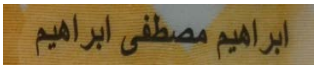


Figure 2: Example of the Arabic Script

TABLE I
THE DIFFERENT FORMS OF ARABIC ALPHABETS

Character Name		Isolated	Initial	Middle	Final
Alif	ألف	ا	ا	ـا	ـا
Ba'	باء	ب	بـ	ـبـ	ـب
Ta'	تاء	ت	تـ	ـتـ	ـت
Tha'	ثاء	ث	ثـ	ـثـ	ـث
Jeem	جيم	ج	جـ	ـجـ	ـج
H'a'	حاء	ح	حـ	ـحـ	ـح
Kha'	خاء	خ	خـ	ـخـ	ـخ
Dal	دال	د	د	ـد	ـد
Thal	ذال	ذ	ذ	ـذ	ـذ
Rai	راي	ر	ر	ـر	ـر
Zai	زاي	ز	ز	ـز	ـز
Seen	سين	س	سـ	ـسـ	ـس
Sheen	شين	ش	شـ	ـشـ	ـش
Sad	صاد	ص	صـ	ـصـ	ـص
Dhad	ضاد	ض	ضـ	ـضـ	ـض
Tta'	طاء	ط	طـ	ـطـ	ـط
Dha'	ظاء	ظ	ظـ	ـظـ	ـظ
A'in	عين	ع	عـ	ـعـ	ـع
Ghain	غين	غ	غـ	ـغـ	ـغ
Fa'	فاء	ف	فـ	ـفـ	ـف
Qaf	قاف	ق	قـ	ـقـ	ـق
Kaf	كاف	ك	كـ	ـكـ	ـك
Lam	لام	ل	لـ	ـلـ	ـل
Meem	ميم	م	مـ	ـمـ	ـم
Noon	نون	ن	نـ	ـنـ	ـن
Ha'	هاء	ه	هـ	ـهـ	ـه
Waw	واو	و	و	ـو	ـو
Ya'	ياء	ي	يـ	ـيـ	ـي

2) The Dotted Challenge: Dotting is widely used to distinguish letters sharing similar graphemes. Fig.3. shows example sets of dotting/un-dotting graphemes, it is clear that the digital variations between the elements of the same set are small. Whether the dots are separated before the recognition process, or recognition features are produced from the dotted script, dotting is a significant source of confusion, and recognition errors in Arabic OCR systems can be occurred.

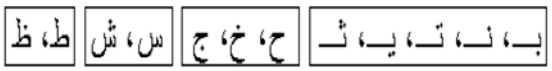


Figure 3: Example sets of dotting-differentiated graphemes

3) Multiple Grapheme Cases Challenge : Due to the variety of connectivity in Arabic orthography; the same grapheme representing the same character can have many shape according to its position within the Arabic word segment (opening, Middle, Ending, Separate) as exemplified by the 4 variants of the Arabic letter "ع" shown in bold in Fig. 4.

Figure 4: Grapheme "ع" in its 4 Positions

4) Character's Size Variation Challenge: The Arabic graphemes have variable height and/or variable width. Furthermore, different nominal sizes - of the same font- do not scale linearly with their actual line heights. At this point, many challenges have been illustrated in the Arabic script characteristics, in the next section; the algorithm of the proposed OCR system will be introduced.

3 SYSTEM OVERVIEW

The data acquisition system consists of a simple wooden structure scanning model. This model has a special place for the mobile and another one for the ID card, the image is captured with a 5 megapixel resolution camera of an Android mobile and the captured images' resolution is 72 dpi.

Using this simple scanning process, images of front and back sides of the ID card are captured by mobile, and then the captured images are sent automatically to the computer by Android service installed in such mobile. The captured image is then passed and processed via the system's software (Matlab library) and the results will be outputted through simple Graphical User Interface (GUI) which contains all the recognized available ID data.

The following sections discuss the methodology of the proposed OCR system in details. Fig.5 shows simple block diagram of the proposed system. Next subsections will explain the related algorithm and the main steps,

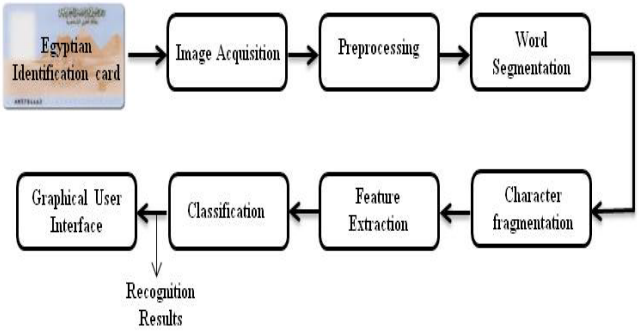


Figure 5: The Proposed Arabic OCR Block Diagram

A. OCR Algorithm

The proposed OCR system consists of 4 main steps which will be introduced within the following subsections.

1) Image Acquisition: As illustrated before, we capture an image of the ID card with an android mobile with 5 Megapixel resolution camera, the captured image is 72 ppi (pixel per inch).

2) Preprocessing and Edge Detection: Preprocessing is a very important and main step in image processing, as scanned images are usually displayed in gray scale or color and also they suffer from noise, varying spaces between letters, varying spaces between lines and other image problems.

Noise removal and edge detection are the two most important steps in processing any digital images to improve the information in the picture so that it can be easily understood by man and to make it suitable and readable for any machine working on those images. So, some preprocessing steps are performed on the image.

Firstly, in the step of thresholding and binarization, the gray image is converted to a 'binary' image. We mean by binary that the image is presented by black and white pixels only. This helps us to work more efficiently on the image than in the gray or the colored form. Binarization is achieved through the process of thresholding, in which a 'threshold' value is chosen and any pixel with a value greater (or less) than this value is converted to a text (or background) pixel. That is, its value is made either 0 or 255.

Noise is the unnecessary information that exists in the image which may has been inadvertently introduced. This may occur because of inefficient input devices used. To remove the noise which may affect the performance of the system, filters are used. So firstly, the "median filter" is used as an example of a non-linear spatial filter, (recall that the median of a set is the middle value when they are sorted), which is used to remove the salt and pepper noise from the card image. Finally, remove the undesired borders by clipping it. Edge detection is a very important step because the edge is one the important and basic features of an image. If the edges of an image are identified correctly, some essential properties such as region, border and outline can be calculated. Edge detection can play a meaning role in different fields such as image segmentation, recognition and image analysis. Many classical edge detectors have been developed over time. However, classical edge detectors usually fail to handle images with strong noises [7], [10], [11].

Mathematical theory and morphology can be used to process and analyze such images. In the morphology theory, images are treated as sets, and morphological transformations are employed to extract features of images [9]. Then, morphological filters are applied to enhance image quality in order to extract useful information about the geometrical structure and accomplish the goal of preserving features while removing noise. So, this algorithm will be applied in the phase of card recognition.

The structuring element is considered to be the building block of the dilation (expanding features) and erosion (shrinking features) processes and, by the way, it is the mainly constructing element of the morphological image processing, it is symbolized by matrix of 0s and 1s, sometimes it is convenient to show only the 1's. The origin of the structuring element must be identified. It could have any shape but its size must be smaller than the original image's size. Fig.6. shows some shapes of structuring elements [12] [13].

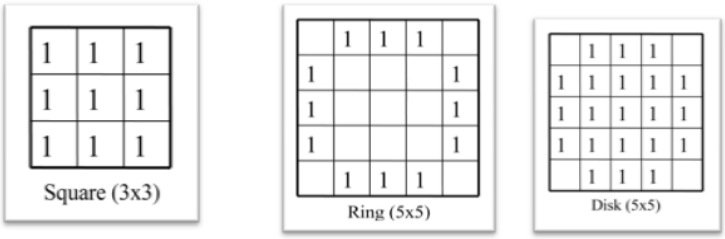


Figure 6: Some shapes of structuring elements

In the proposed system, the rectangle and line (with 0 & 90 degrees) structuring elements are used as they are more suitable for our dilation and erosion operations, see Fig. 7.

Figure 7: Line and rectangle structuring elements

Dilation is processes that produce objects in binary images controlled by a form referred to as a structuring component. Dilation of original image (A) by structuring element (B) means that for each point x belongs to B, we translate A by those coordinates, and then we take union of these translations, as shown in the following equation [15].

$$A \oplus B = \{z|(\hat{B})_z \cap A \neq \emptyset\}$$

The above equation denotes the dilation of A by B is the set consisting of all the structuring part source locations where the reflected and translated B overlaps at least some portion of A. Such translation of the structuring element in dilation is similar to the mechanics of spatial convolution.

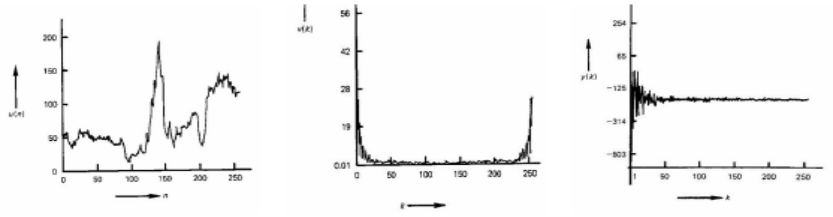
Erosion is an operation that thins or shrinks objects in binary images controlled by a shape referred to as a structuring element. Erosion of original image (A) by structuring element (B) means that the output image has a value of '1' at each location of the origin of (B), such that the element only overlaps 1-valued pixels of (A) [15].

$$A \ominus B = \{z|(B)_z \cap A^c \neq \emptyset\}$$

The above equation denotes that the erosion of A by B is the set of all structuring element origin locations where the translated B has no overlap with the background of A.

3) Feature Extraction Module: Features extraction is very important feature in accomplishing good recognition performance in pattern recognition systems. It has been defined as the process of extracting information that is mostly useful for the purpose of classification from the raw data, as it involves simplifying the amount of resources required to describe a large set of data accurately, so it is considered to be a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant, the input data can be transformed into a reduced representation set of features (named feature vector). So transforming the input data into set of features is called the feature extraction.

Discrete Cosine Transform (DCT) is one of the most popular techniques used in feature extraction [9], [10], [14]. In particular, a DCT is a Fourier-related transform like to Discrete Fourier Transform (DFT), but DCT is more capable in data reduction as it stores nearly all of image details in few coefficients as shown in Fig. 8.



(a) The Source (b) DFT (c) DCT

Figure 8: Energy distributions in different transforms

The proposed solution uses mathematical tool of the Discrete Cosine Transform DCT. Such DCT-based feature extractor is used in image processing, especially for lossy data compression, because it is capable of packing the energy of spatial sequence into few coefficients as possible so it has strong energy compaction property. So, a larger number of coefficients get wiped and great bit savings for the same loss.

Two Dimensional -DCT is applied to the whole image [9]-[10], then most of signal information tends to be concentrated in a few low-frequency components and approximately most of the important data and details of the image are then allocated at the upper left corner of the image, as shown in Fig.9.

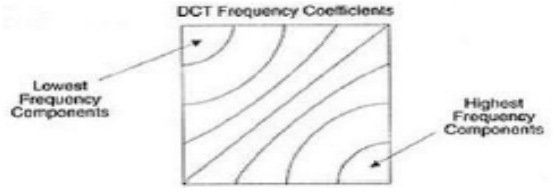


Figure 9: The frequency distribution of two dimensional- DCT

Eye is most sensitive to low frequency components (upper left corner), so Zig-Zag scanning method is used to group low frequency coefficients in top of a vector with a certain technique, as shown in Fig. 10.
After applying Zigzag scan, a vector of 20 elements is created for each model in the training set; this vector is the feature vector that will be used later in the next stage of classification and decision making.

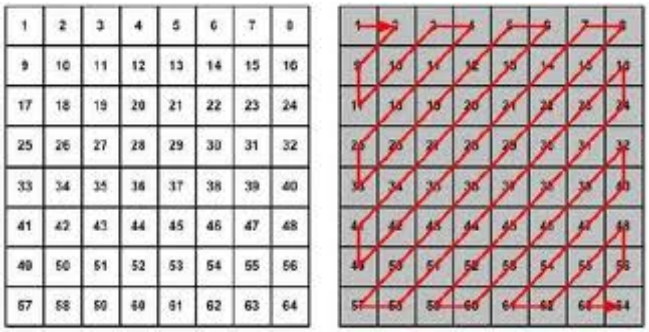


Figure 10: Zig - Zag Scan

4) Classification and Decision Making: The main role of the classifier is to compare the feature vector of each block of data segmented from ID card with the previously built model of the training set, and then provide the system with information about the nearest neighbor to this block of data and the Euclidean distance between them. The Euclidean distance between any 2 vectors (q and p- each one of them has (n) elements) can be calculated according to the following formula:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} (q(i) - p(i))^2}$$

where, “q” and “p” are two feature vectors with size (n) equals 20 elements.

B.ID Recognition Algorithm Steps

First, the preprocessing steps are applied on the whole ID card except the erosion step, after performing dialation or filling operation the result will be displayed as shown in Fig. 11.

Figure 11: ID card after Dilation Process

Then, the wanted data is detected and then the areas of the data are segmented with rectangular frames, each type of data is sent to its specified function.
There are two types of algorithms depending on the data types, an algorithm for numbers like ID National number, release date and expiry date of the ID cards and the other for words which is specified for all other types of data.
The data types are (First Name & Last Name – Address – ID National number – Job – Job Address – Religion & Type & Marital status – Husband name – Expiry date& Release date). A training data set, containing all forms of Arabic language alphabets, numbers from (0-9) and some characters like (/ , -), is passed to the words, function and number's function respectively.

1) Numbers Function's Algorithm: A training data set containing all numbers from (0-9) is passed to this function, and also some characters like (/, -).
Firstly, we get the feature vectors for the numbers (0-9) as a data set, and then we get the feature vectors for the unknown national ID card number which we want to detect its 14 numbers, Fig.12 shows an ID card number example.

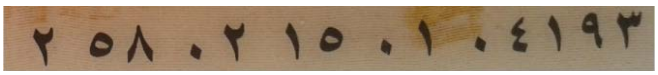


Figure 12: ID Card Number Example

The following steps will be used to perform preprocessing:
Thresholding and binarization: In this step, the RGB ID number image is converted to a 'gray' image, then to 'binary' one.
Noise filtering: the "median filter" is used as an example of a non-linear spatial filter, to remove the salt &pepper noise from the image number.
Mathematical morphology (MM): which consists of two main basic operations; Dilation and Erosion, the line (with 0 & 90 degrees) structuring element is more suitable for dilation and erosion operations. Fig. 13 shows ID card number after performing preprocessing steps.

Figure 13: ID card number after performing preprocessing steps

Segmentation: Then, by using some functions, the ID number image is divided into 14 segments (regions). Then, for each region (number) we get its properties (area, centriod, bounding box) and by using the bounding box vector, which contains [X_{min}, Y_{min}, width, height] we could detect its boundaries and the image is segmented and be ready for feature extraction stage Fig. 14 shows ID card numbers after segmentation.

Figure 14: Segmented ID card numbers

For each region, which represents one number in the ID number, we get the feature vector and by using our classifier, we get the location of the nearest number to database, also we get the minimum distance between the region feature vector and the other feature vectors in database. According to this minimum Euclidean distance, we can take the right decision. Finally, we have a vector containing the 14 numbers of the National ID Number.

2) Words Function's Algorithm: In these functions, the training data set is containing all forms of Arabic language graphemes. The proposed algorithm for segmentation and detection of Arabic words is starting to get the feature vectors for each Arabic character from the database. Then, three steps are applied to perform preprocessing: (1) Thresholding and binarization; (2) Noise filtering; and (3) Math morphology (Dilation and Erosion).
Deskewing to fix the wrongly rotated words: to remove the undesired rotation in the image, rotate the image with many angles, for each angle apply the horizontal projection profiles and find the highest peak, then choose the rotation angle equals to the angle which has the highest peak and rotate the image by it. Deskewing mechanism is showing good performance with the fields that have many words, but showing bad performance with the single word, so for the fields of data that have just one word, like first name field, the deskewing mechanism is deactivated, but for sentences like address or last name, it's useful to activate this mechanism.
After the cut area is sent, words segmentation and separation are done by noticing the intra-spaces between words of the sentence, by experiment the number of zeros (space) between words is determined to be 25 zero minimum. Sentence image is vertically projected, by applying vertical projection on image, the vertical projection profile is defined as:
$P(j) = \sum image(i, j)$, where p(j) is the vertical projection of the image for column j and the image(i, j) is the pixel value at (i, j). See Fig. 15.

Figure 15-a: The image after filling and deskewing

Figure15-b:The image after erosion and deskewing

By using this vertical projection, the spaces in the sentence can be located by applying a certain condition on the length of these spaces, so the location of spaces between words in the sentence can be easily recognized, the regions of words are detected and we got segmented words and the words are separated, see Fig.16.

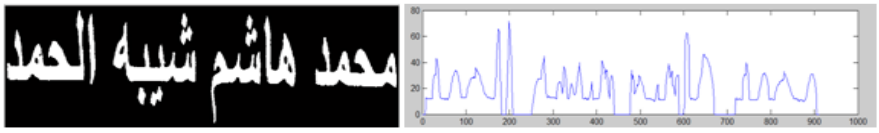


Figure 16: The words after segmentation

Detecting baseline is one of the main majorities in the preprocessing step of Arabic OCR system, as it can be used for both skew normalization and segmenting the text into words or characters. The baseline detection is very important in Arabic OCR, because it can be used to segment the Arabic text to characters and make the text ready for the feature extraction stage. Also, baseline has been used by most of the OCR systems [12]-[14].

The horizontal projection method is used by the OCR researchers to detect Arabic baseline, and it works well with the printed text. This method detects the Arabic baseline by reducing the 2D of data to 1D based on the pixels of the text image, and the longest peak after projection will define the baseline range.

The horizontal projection profile is defined as: $P(i)=\sum \text{image}(i,j)$,where p(i) is the horizontal projection of the image for row i, and the image(i , j) is the pixel value at (i , j).

Determine the baseline: that depends on the iteration with angle to detect Arabic baseline with the horizontal projection, the highest peak corresponding to our rotation angle is used to determine the baseline, and Fig. 17 shows text after detecting the baseline.

Figure 17: Text after detecting the baseline

For fragmentation of words to characters, we apply vertical projection to the word to detect the beginning and end of each character in the word, as shown in Fig. 18.

Figure 18: Character Fragmentation

The classification step starts with the feature vector of the unknown segmented character, by applying the proposed classifier using the Euclidian distance between the unknown character feature vector and all training feature vectors. According to the minimum distance, we can know the nearest character and take the right decision.

3) Gender, Religion and Marital status Function's Algorithm: To increase the system accuracy, we make a separate function for the standard words in the national ID card like gender, religion, and marital status.
Firstly, we get the feature vectors for these standard words as a data base. For religion ('مسلم','مسلمة','مسيحي','مسيحية')'. For marital status ('أعزب ',' انسة',' متزوج','متزوجة','مطلق','مطلقة',' ارمل',' ارملة ') and for gender ('انثي ',' ذ كر'). Then, for the unknown word which we want to know, we apply two steps to perform preprocessing [12] [13].
Thresholding and binarization: In this step, the RGB image is converted to a 'gray' image, then to 'binary' one. Noise filtering: we used the "median filter" as an example of a non-linear spatial filter to remove the salt &pepper noise from the card. Then, we apply the mathematical morphology (MM): which consists of two main basic operations; Dilation and Erosion, we used the line (with 0 & 90 degrees) structuring element as they are more suitable for our dilation and erosion operations.
For the decision step, we get the feature vector of the unknown segmented area, by using the proposed classifier we get the minimum Euclidean distance between this feature vector and the feature vectors in the standard words database. According to this minimum Euclidean distance, the right decision can be taken.

4 EXPERIMENTAL WORK AND RESULTS

This section presents the results obtained, and discusses the limitations and problems which affect the accuracy of the system, and also discusses solutions for some of them.
The number of collected ID cards is 40; images are captured by a 5 megapixels camera which takes pictures with resolution 72 ppi. We are ensured to be taken at random and in different environments, as the changes in brightness or

source of light or the homogeneity will affect the accuracy, the ideal situation of scanner solves this problem as it provides a homogeneous, fixed distribution of light for all parts of the ID card. For the proposed system, sunny environment provides this homogeneous distribution, so it's chosen to be the default environment for the design, and as we'll see in the following sections, the change of environment is a reason of some bad accuracy, the 40 ID cards are divided into 10 ID cards for training phase and 30 ID cards for testing phase. Table II shows the accuracy of the system for the testing phase before doing corrections.

TABLE II
SYSTEM ACCURACY BEFORE CORRECTION

Phase	System Accuracy		
	Total Number	Correct Number	Percentage
Segmentation	374	360	96.26%
Numbers	1249	1246	99.76%
Words	726	675	92.97%

The average time of run for a single card reaches 16 seconds, and the maximum time of run for a single card is 22 seconds. The following sections will discuss some reasons that result in errors at segmentation and words recognition phases, and then we will suggest some solutions to increase system's accuracy.

A. Problems of Segmentation Phase

Segmentation phase is the process of cutting out the important information from the ID card after finishing the preprocessing phase. In case of ID card where the locations of data are almost fixed, so it can be defined easily if segmented information is the first name, last name … etc. But, in some ID cards there are some unusual situations causing errors in segmentation process. The following subsections will illustrate some of these problems noticed from our experimental testing.

1) Wrong locations of information problem: As previously illustrated, the location of information is almost fixed in different ID cards. Consequently, every segmented data can be defined from its location. But as displayed in Fig. 19-a, the information in the back of ID card in figure (19-b) is shifted obviously up than any other normal ID card and we can notice that by comparing the location of information in both cards. This problem makes the system detects the job address field in ID card in Fig. (19-b) in the field of job and the religion in the field of job address, and can't define the job, marital status and husband name information, as shown in Fig.19-c.

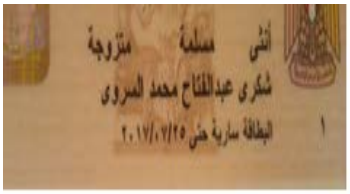
(a) (b)

Job :	مدرسة الابكانية
Job address :	معلم
Gender :	أنثى
Religion :	not found
Marital state :	not found
Husband name :	not found

(c)

Figure 19: Results of wrong locations of information

Another problem of wrong located information, cards which don't have expiry date information field for married females, husband name is shifted down and the system segments it as expiry date which causes an error of segmentation .Fig. 20-a shows an ID card with normal husband name location but Fig. 20-b shows a wrong location of husband name in other card.



(a)

(b)

Figure 20: Wrong husband name location

2) Dilation Process problem : Dilation process is used to fill the spaces between words, so the locations of text can be defined and then the system can define the type of this information, but in some cases, some information fields could be under dilated or over dilated. Under dilation problem would cause partially successful segmentation, and over dilation problem is linking unnecessary information to a necessary one, so the system will be unable to detect the necessary information and considers it missing information.

3) Noisy Information problem: If there is noisy information in the ID card caused by an error in preprocessing or resulted from under dilated information, as previously illustrated, this noisy information is detected as necessary information which causes missing of the needed information.

Table III highlights 374 scanned cards to be segmented, the system successfully segmented 360 fields of information and failed in 14, Table III shows the error rate of each problem previously illustrated.

TABLE III
ERROR RATE OF SEGMENTATION PHASE PROBLEM

	Wrong locations problem	**Dilation process problem**	**Noisy information**	**Other**
Error Rate	10	2	1	1

B. Problems of Words Recognition Phase

Words recognition process is the process of recognizing characters in an image and translating them into computer textual format by applying the recognition mechanism on the image, various errors happen in the experimental work. The following subsections discuss some common errors noticed in testing phase.

1) Baseline Detection problem: The first step of baseline detection process is getting the horizontal projection of word image, defining the row that has maximum value in the projection, then defining a range of rows around the row of the maximum value. But, in some testing words, this range is not enough to detect the baseline accurately, and this happens because of different sizes of words in the same field for different ID cards or because of the skewing of the image was not fixed at the deskewing process. Failing in detecting the baseline accurately would cause failing in characters segmentation, as the characters get linked together, as shown in Fig. 21, this is a wrong recognized word due to wrong detection of baseline.

Figure 21: Baseline detection problem

Another reason of linking characters is the Arabic character 'ع ' when it appears as the first character of the word, in some cases it sticks with the following character because of the very small space between them as can be seen in Fig. 22, so, the baseline detection process can't separate them accurately.

Figure 22: Stuck characters problem

2) Over and Under Segmentation problem: In character segmentation process, when words have different sizes rather than the default size in a certain field, this leads to wrong segmentation for some characters like the Arabic character 'س' for example. Over-segmentation occurs when single character is segmented to more than one segment, and under-segmentation occurs when more than one character are considered as one character, Fig. 23 illustrates the two cases.

Figure 23: Over and under segmentation

3) Over Erosion: As previously mentioned that for the same field of information, e.g. job field, the words could differ in size for different ID cards, and as the erosion process is fixed for all words in the same field, over-erosion could happen to the smaller size words, it results in errors in the recognition process. One of the common problems caused by over-erosion is errors occurred in the word which has the Arabic character 'ء/hamza', over-erosion is affecting this character as the eroded 'hamza' is recognized by the system to be a dot. Such problem represents a word has over eroded 'hamza', the system recognized it as another Arabic character 'ـٮ/noon' in the recognition process. Fig. 24 shows an over erosion example.

Figure 24: Over erosion example

4) Environment Effect on the System Accuracy: It's mentioned before that taking pictures for the ID cards in different environmental conditions affects the system accuracy, as the image of the ID card in a certain environment could be clearer than another image for the same ID card in different environment, and it should be considered that the testing ID cards' pictures are taken in different environments to test the capability of the system to deal with different circumstances, and it's shown that the more clearer picture with homogeneous light distribution the more accuracy is obtained. Fig. 25 shows the response of the system to two pictures for the same ID card but in different environmental conditions.

Figure 25: Same ID card in two environmental conditions

C. Suggested Correction

As we represented the problem of wrong location of husband name information in most ID cards of females which don't have expiry date information field, a correction for this problem is suggested that depends on linking the information fields with each other, by taking the advantage that the functions of gender, religion and marital status are perfectly recognized in all tested ID cards because they are depending on correlation method of recognition. So, if the system detected that the ID card is for a married female, so the husband name should be found, and if it's not found the system checks the output of expiry date field recognition function. If the output is not logical enough, the system refuses this output and sends the information of expiry date field to the function of husband name recognition. Fig. 26 shows the output of an ID card that has this problem before and after the correction.

Figure 26: Correction effect

This algorithm solved 4 errors out of the 14 errors in the segmentation phase, and to ease following-up the algorithm, we'll write it in steps. Run the functions of gender and marital status recognition. If the card is for married female, check the husband name. If there is no information in husband name field, run the function of expiry date recognition directly.
If the output of expiry date function is not logical enough, refuse this output and send the information in expiry date field to the function of husband name recognition. If the output is logical, accept this output as expiry date and inform that system didn't find husband name.

D. Final Results

The correction enhances the accuracy of segmentation phase, as it solves most of the wrong location cases of husband name, the final results of testing 30 ID cards are shown in Table IV.

TABLE IV
SYSTEM ACCURACY AFTER CORRECTION

Phase	**System accuracy (after correction)**		
	Total number	**Correct number**	**Percentage**
Segmentation	374	364	97.33%
Numbers	1249	1246	99.76%
Words	743	688	92.60%

5 CONCLUSIONS AND FUTURE WORK

To summarize, the proposed Arabic OCR system is used to extract data from the image of Egyptian ID cards and then it recognizes this data and translates it into computer textual format. This system could be used in creating database for a lot of ID cards easily and fast or in verification between data of ID card and previously saved database. The evaluation of the presented system is excellent in data segmentation and in numbers recognition and it is very good in recognizing characters due to the challenges facing the system in this phase.

As was discussed before, the proposed system is very helpful for business checking and saving personal ID's information. Also, it saves valuable time needed to type these data manually and also gives a very good recognition accuracy, which makes the system reliable enough to be applicable in governmental authorities or even companies. There are many applications that could need to use this system, like election system, Wallet services, Banking card, security services …etc.
The proposed system solution could be enhanced in several ways, the first way is to make better design for the process of taking shots by camera to provide the system with fixed source of light and with fixed intensity in any environment, or by using a scanner provided by motor to satisfy fast processes. This way of development will end most of the common problems facing our system which are changing source of light and intensity of light on all parts of the ID card.
The second way is to test new methods of character segmentation with, or even without, our method which is depending on baseline detection, like using HMM in character segmentation or any other methods, in order to achieve higher accuracies of character recognition.

The third way of development is to develop the system to be used for other purposes and be more generalized in the field of recognizing documents which have fixed locations of information, like ID cards, license cards and passports, which will make the system suitable for a lot of governmental institutions and authorities.

The fourth way of development is to make the system deal with any printed documents with non-fixed information locations and different font sizes and types, like business cards and OCR systems for books and papers. The difficulty of this development is the existence of many font types which should be considered in the training phase, so the training phase will be very complex. Then as an extra development, making android application that allows the businessman to take shots with his mobile camera for any business card, and the system extracts data and saves it in the database in the mobile and when the mobile connects with the personal computer of the businessman, he can export and import data to and from the database.

Finally, the fifth way of development is to develop the system to deal with handwritten documents, and this development needs a huge training set to train the system to recognize handwritten sentences, and this system could be used with forms in any business company or governmental authority.

ACKNOLEDGMENT

The teamwork of the ***"Arabic Printed OCR System"*** project was funded and supported; by the NSTIP strategic technologies program in the Kingdom of Saudi Arabia- project no. (11-INF-1997-03). In addition, the authors acknowledge with thanks Science and Technology Unit, King Abdulaziz University for technical support.

REFERENCES

[1] A. Cheung, M. Bennamoun, N.W. Bergmann, "*An Arabic optical character recognition system using recognition-based segmentation*", Pattern Recognition, Volume 34, Issue 2, February 2001, Pages 215–233.

[2] M. S. Khorsheed, "Off-Line Arabic Character Recognition – A Review", Pattern Analysis & Applications (2002) 5: pp. 31–45.

[3] M. Rashwan, M..Fakhr, M. Attia and M. EL-Mahallawy, "*Arabic OCR System Analogous to HMM-Based ASR systems* ", http://www.rdi-eg.com/Intro_to_NLP/Paper3.pdf 2007.

[4] A. Mesleh, A. Sharadqh, J. Al-Azzeh, M. Abu-Zaher, N. Al-Zabin, T. Jaber, A. Odeh and M. Hasan, "*An Optical Character Recognition*", Contemporary Engineering Sciences, Vol. 5, 2012, no. 11,Pages 521 – 529, http://www.m-hikari.com/ ces/ces2012/ces9-12-2012/meslehCES9-12-2012.pdf

[5] A. Zidouri, M. Sarfraz, S . A. Shahab and S . M. Jaf ,"*Adaptive Dissection based segmentation of printed Arabic text* ", Information Visualization, 2005, Ninth international conference on information visualization.

[6] R. C. Gonzalez, R. E. Woods and S. L. Eddins, "*Intensity Transformation and Spatial Filtering*," in Digital Image Processing Using Matlab", 2009 by Gatesmark, LLC, pp.68–69.

[7] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*. 3rd. Chapter 5, Prentice Hall, 2009.

[8] E. Dougherty, "*Mathematical Morphology in image processing*", 2003, http://ebooks.spiedigitallibrary.org/book.aspx?bookid=159

[9] K. SaiCharan, "*A Block DCT based Printed Character Recognition System*," March, 2006 http://www.cs.ucr.edu/~scharan/assets/Optical.Character.RecoRecogni.pdf.

[10] J. Vasavada, and S. Tiwari,"*A Hybrid Method for Detection of Edges in Grayscale Images*", Image, Graphics and Signal Processing, 2013, 9, 21-28.

[11] K. SaiCharan, "*A Block DCT based Printed Character Recognition System*", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-2, Issue-6, May 2013, http://www.ijitee.org/attachments/File/v2i6/F0820052613.pdf.

[12] M. Zeki Khedhr,G. Abandah,"*Arabic Character Recognition using Approximate Stroke Sequence*", (LREC 2002), http://gandalf.aksis.uib.no/ lrec2002/pdf/ws13/Khedher.pdf.

[13] M. Al-A'ali and J. Ahmad, "*Optical Character Recognition System for Arabic Text Using Cursive Multi-Directional Approach*", Journal of Computer Science, Volume 3, Issue 7, 2007.

[14] H. Imtiaz and A. Fattah, "*A DCT-based Feature Extraction Algorithm for Palm-print Recognition*", Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on information visualization, Pages 657 – 660.

[15] Z. Ren ; Z. Fan ; P. Ran ; J. Li, "Straight interference fringes thinning algorithm", 5th International Symposium on Advanced Optical Manufacturing and Testing Technologies: Optical Test and Measurement Technology and Equipment Yudong Zhang; José Sasián; Libin Xiang; Sandy To Dalian, China, April 26, 2010 Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference.

BIOGRAPHY

Mohsen A. A. Rashwan, Prof. of Electronics & Communications Department, Faculty of Engineering, Cairo University, and Chairman & MD of RDI. He received a Ph.D. degree in Electrical and Computer Engineering from University of Queen's University, Kingston, Ontario, Canada. Before, he received his Master Degree of Electronics and Electrical Communications, *Digital Filters Implementation using Microprocessors*, Faculty of Engineering, Cairo University, Egypt. Over 100 papers are published in international proceedings and conferences. (See some relevant papers in the end of this CV). Over 80 theses under his supervision (finished and current): 34 Ph.D.'s and 47 M.Sc. and 4 Master of Arts. Many of my ex-post graduate (MSc. & Ph.D.) students are currently recruited in the core of top world hi-tech companies and research centers like IBM-WRC, Lucent Technologies, Microsoft, etc. Over 350 graduation projects are realized under his supervision.

Hassanin M. Al-Barhamtoshy received the B.S. degree in electronic and communication engineering from Cairo University, in 1978, and the M.S. degree in systems and computers engineering from the Al-Azhar University, Cairo, 1985. In 1992, he received the Ph.D. degree in systems and computers engineering from Al-Azhar University, Cairo. During 1992–1997, he was an Assistant Professor in the Department of Systems and Computer Engineering at Al-Azhar University. During 1996-1997 he was an Assistant Professor in Computer Science at KAU University, Jeddah, Saudi Arabia. During 1998-2002 he was an Associate Professor in Computer Science at KAU University, Jeddah, Saudi Arabia. He is currently Professor in the Department of Computer Science and Information Technology at Faculty of Computing and Information Technology, KAU University (2003-now). His research interests include language processing and machine translation, image processing, software engineering, intelligent systems, speech processing, e-learning, and RFID.

التعرف على بطاقات الهوية بنظام التعرف على حروف اللغة العربية

[1]أميرة عبد الكريم شعبان، [2] أشرف حسين عبد الرسول، [3] إسراء شكري عبد الفتاح، [4] علا علاء الدين عبد النعيم، [5] محسن رشوان
قسم هندسة الإلكترونيات و الاتصالات الكهربية، كلية الهندسة، جامعة القاهرة
[1] amira_shaban22@yahoo.com
[2] h_ashraf16@hotmail.com
[3] engesraa_2013@hotmail.com
[4] olaalaa_2013@yahoo.com
[5] mrashwan@rdi-eg.com

حسنين محمد البرهمتوشى
كلية الحاسبات وتقنية المعلومات، جامعة الملك عبد العزيز، جدة، المملكة العربية السعودية

hassanin@kau.edu.sa

المستخلص

إن التعرف الضوئي على الحروف و خاصة حروف اللغة العربية لهو مجال بحثي صعب و مليء بالتحديات و لكنه مهم للغاية لأنه يدخل في العديد من التطبيقات المهمة التي تحتاج للتعرف على نص مستند ما أوتوماتيكيا من صورته.

هذا البحث يعرض مشروع نظام قارئ لبطاقات تحديد الهوية المصرية و الذي يقوم باستخراج المعلومات المهمة و الأساسية من صورة البطاقة الملتقطة و التعرف عليها ثم تحويلها إلى قاعدة بيانات سهلة الادراج في الحاسب.

هناك العديد من التطبيقات لهذا النظام فهو من الممكن أن يستخدم في تسهيل اجراءات العملية الانتخابية؛ حيث يكون من السهل التحقق من بطاقات المواطنين أوتوماتيكيا و التأكد من الادلاء بأصواتهم و بذلك يمنع حدوث تزوير أو ما شابه .

أيضا من الممكن استخدام هذا النظام في المنشآت الحكومية و المصانع التي تحتاج إلى قواعد بيانات لعديد من المواطنين و التحقق من هويتهم. ويستغرق استدعاء مكتبة هذا النظام 16 ثانية للتعرف على بطاقة واحدة و هذا ما يوضح مدى واقعية تطبيقه و استخدام هذا النظام في الحياة اليومية و من المتوقع للنظام أن يؤدي بشكل أفضل حين الانتقال من المرحلة الاكاديمية إلى مرحلة تصنيع المنتج النهائى .