Journal of Theoretical and Applied Information Technology
15th October 2017. Vol.95. No 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ON ARABIC OBJECT CHARACTER RECOGNITION USING DYNAMIC TIME WARPING

[1]ABDELWADOOD MESLEH, [2]OMAR ARABEYYAT, [3]SHARHABEEL ALNABELSI, [4]JAMAL AL-NABULSI

[1,3]Computer Engineering Department, Faculty of Engineering Technology, Al-Balqa Applied University, Amman, Jordan
[2]Computer Engineering Department, Faculty of Engineering, Al-Balqa Applied University, Salt, Jordan
[4]Department of Medical Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan.
[4]Biomedical Engineering Department, Faculty of Engineering, The Hashemite University, Zarqa, Jordan.
E-mail: [1]wadood@fet.edu.jo, [2]arabiat@bau.edu.jo, [3]alnabsh1@bau.edu.jo, [4]j.nabulsi@ammanu.edu.jo, [4]jian@hu.edu.jo

ABSTRACT

Due to the large volume of Arabic texts in many generated and historical documents, it is essential to use computers in order to make generated texts editable, this is actually the main task of Arabic Object Character Recognition (OCR) systems. The task of automatically OCRing is to type documents within close-to-human performance, such OCR system is still an open research problem. In this paper, we propose an Arabic OCR based on Dynamic Time Warping (DTW) algorithm that is empowered to properly recognize Arabic words. Rather than using the usual practice of character segmentation, this paper proposes a segmentation of Arabic texts into lines and characters. The proposed Arabic OCR algorithm overlaps the segmentation and the recognition processes – an online segmentation-recognition. That is, in order to overcome the challenges of segmenting highly cursive Arabic texts into isolated characters. The accuracy of the proposed Arabic OCR algorithm is tested on randomly selected articles from Jordanian newspapers. Interestingly, results demonstrate the robustness of our proposed Arabic OCR algorithm that achieves 96.2% character recognition accuracy in the worst case.

Keywords: *Object Character Recognition, Dynamic Time Warping, Online Arabic OCR, Typed Arabic OCR*

1. INTRODUCTION

OCR [1,2,3,4,5] is a field of research in artificial intelligence, OCR scans printed or handwritten texts, performs character analysis on the resulting images, and finally extracts the corresponding text which can be stored and manipulated electronically as any standard electronic document. Despite of many research work and development, OCR modules are still nowhere near human's reading capabilities. Many of existing OCR modules are not always perfect, such imperfectness has an adverse effect on the efficiency of the many OCR applications, such as information retrieval systems.

OCR systems are categorized into two types [5]: offline or online. First, offline OCR systems are further categorized into handwritten and typed OCRs. Second, in online OCR systems, characters and words are recognized as soon as they are written, such as writing with a special pen on tablets and smart phones. Number, order, and direction of strokes assist the recognition process in online OCRs. The main advantage of online OCRs over offline OCRs is its interactive nature; however, the main disadvantage is the necessity of dynamically processing documents. Currently, handwritten texts on modern smart phones are efficiently converted to texts using online OCR systems; however, there is a lot of room for research and development for improving Arabic OCR that suffers a lag in technology. In Offline OCR systems, the recognizer scans a static representation of typed or handwritten document, analyses it and finally converts it to editable texts. The main challenges of offline OCR systems are due to the variety of handwriting or printing styles, especially for Arabic language. There are many applications for offline OCR systems, such as vehicle plate recognition [1,2,3].

An OCR system starts with an image that may be retrieved by a scanner or a digital camera. Then,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

during the pre-processing phase, text is extracted and segmented. After that, in recognition phase, features of the segmented components are extracted and finally a classifier is implemented to extract editable text characters. Typical pre-processing includes [1,2,3]:

- ✓ Noise removal using a low-pass filter, binarization, histogram and morphology based methods.
- ✓ Binarization that converts grey images to binary images using thresholding techniques.
- ✓ Skew detection and correction, in order to estimate the skew angle of text lines in the text image.
- ✓ Page analysis that allows to identify and to categorize regions of interest in the text image.
- ✓ Segmentation of the characters may affect the recognition performance, as the output of the segmentation process is the input of the recognition engine. Notice that some OCR approaches skip the segmentation process.
- ✓ Feature extraction: the segmented characters are fed into a feature extraction method. Feature extraction methods include projection histograms and zoning.
- ✓ Classification: the feature extraction goal is to find the best subset of features that precisely and uniquely defines the shape of the corresponding character. The classification algorithm assigns labels to character images in order to enable the production of editable characters.

In 1960s, OCR research was started and the first commercialized OCR was developed by IBM, in 1970s, Toshiba was developed an OCR system that recognized regular typed and handwritten characters. In 1980s, OCR systems were improved to deal with poor quality documents. Starting from 2000, reliable OCR systems were improved to operate on complex and skewed documents that contain text, graphics and mathematical symbols. After more than 50 years of research, developing an optical OCR system with human reading capabilities still unachieved for Arabic characters, as OCR involves many techniques in many areas, such as image processing, pattern recognition, natural language processing and artificial intelligence.

Arabic [6] is a complex script and a cursive-type language which is used by more than one billion users in the world. It is written from right to left, and it consists of 28 characters, each of them has two to four different forms which depend on its position within the Arabic word or sub-word. The Arabic character set and their variations are shown in Table 1. Each Arabic word can have one or more sub-words. On one hand, most characters have dot(s) or zigzag(s) associated with the character and this can be above, below, or inside the character. On the other hand, many of them have similar shapes. Sometimes, the position or number of these secondary strokes is the only difference between some characters. Moreover, some Arabic words may horizontally overlap and characters may stack on others. These characteristics of Arabic characters and words may induce problems for both the word and the character segmentations and may involve many problems in developing an Arabic OCR system. Moreover, the Arabic recognition should occur from top to bottom and from right to left.

The problem of character recognition is twofold, along with the purely mathematical part there is a linguistic one. When solving such kind of problems, collaboration of linguists and computer programmers gives good results in the creation of linguistically oriented computer programs. Nowadays, there is effective OCR software developed for Latin or Cyrillic texts. On the other hand, for Arabic texts, the problem seems to be more complex.

To a certain extent it is caused by the peculiarities of the Arabic script, among which one can mention a big number of morphological and graphic derivatives, cursive stile of writing, which presents a continuous flowing line, joined-up writing of many prepositions, particles, etc. Sometimes the elements of the characters (the dots) may be presented in disperse, and may be located apart from the main element of the letters [7].

Despite the abundance of work and research materials dedicated to the problem of Arabic OCR made up today, the problem is still far from to be resolved. In order to provide accuracy in Arabic text recognition a set of special methods should be used along with the classical attitudes for OCR. Among these, there is a development of the grammatical and lexical analyzers, consideration of such useful data as the relative frequency of use of letters or words, the comparison with existing databases, such as e-dictionaries, software lemmatization [7].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Shapes of Arabic alphabets in different positions.

Name	Isolated	Initial	Medial	Final
Alif	ا	ا	ـا	ا
Baa	ب	بـ	ـبـ	ـب
Taa	ت	تـ	ـتـ	ـت
Thaa	ث	ثـ	ـثـ	ـث
Jiim	ج	جـ	ـجـ	ـج
Haa	ح	حـ	ـحـ	ـح
Khaa	خ	خـ	ـخـ	ـخ
Daal	د	د	ـد	ـد
Dhal	ذ	ذ	ـذ	ـذ
Raa	ر	ر	ـر	ـر
Zaay	ز	ز	ـز	ـز
Siin	س	سـ	ـسـ	ـس
Shin	ش	شـ	ـشـ	ـش
Saad	ص	صـ	ـصـ	ـص
Daad	ض	ضـ	ـضـ	ـض
Taa	ط	طـ	ـطـ	ـط
Dhaa	ظ	ظـ	ـظـ	ـظ
Ayn	ع	عـ	ـعـ	ـع
Ghayn	غ	غـ	ـغـ	ـغ
Faa	ف	فـ	ـفـ	ـف
Qaaf	ق	قـ	ـقـ	ـق
Kaaf	ك	كـ	ـكـ	ـك
Laam	ل	لـ	ـلـ	ـل
Miim	م	مـ	ـمـ	ـم
Nuun	ن	نـ	ـنـ	ـن
Haa	ه	هـ	ـهـ	ـه
Waaw	و	و	ـو	ـو
Yaa	ي	يـ	ـيـ	ـي
Hamza	ء	ئـ	ـئـ	ـئ

DTW [8] is one of the well-known techniques that finds an optimal alignment between two given (time-dependent) sequences under certain restriction, DTW warps the two sequences in a nonlinear fashion to match each other. To stretch (or compress) two-time series in order to make one resemble the other as much as possible, the distance between them computed, after stretching, by summing the distances of individual aligned elements. In OCR system, one of the main advantages of DTW is its ability to recognize, properly, words or characters without prior a separate segmentation step. In 1989, Khemakhem had initially adapted DTW to Arabic OCR [9], in which DTW provided a word recognition rate of 97%, however, the recognition rate was dependent on the size of the text font used.

In [10], authors proposed a distributed Arabic handwriting OCR system based on a parallel Fast DTW algorithm [11] via cloud computing technologies [12] and concluded that cloud computing technologies are beneficial to accelerate the process of Arabic handwritten recognition. Similarly, authors in [13] proved that their distributed computing system implementation is promising to speed the Arabic OCR based on the DTW algorithm. In [14], authors presented an inertial-sensor-based digital pen (inertial pen) associated with DTW-based recognition algorithm for handwriting and gesture recognition, the user-independent recognition rates vary between 87.3% and 98.1%, moreover, their user dependent recognition rates vary between 93.0% and 99.8%.

Rather than using the ordinary character segmentation process [15,16] in which the lines are segmented into lines, the lines are then segmented into words, and finally, the words are segmented into characters. In this paper, the proposed OCR algorithm segments the lines into characters directly; it overlaps the character segmentation and the recognition processes. This choice is to avoid some of the Arabic text challenges and to speed the recognition process by dropping the word segmentation step.

The proposed OCR algorithm aims at recognizing Arabic characters in Jordanian newspapers - small character sizes- with a high character recognition accuracy rates. Moreover, this paper is to implement a parallel version of the whole OCR process to speed the proposed OCR algorithm to be applicable for Arabic newspaper text recognition.

Moreover, this paper compares the recognition character rate using DTW and correlation [17] in the segmentation - recognition step. Noting that authors in [17] presented a simple approach for Arabic OCR that deployed correlation and dynamic-size windowing to segment and to recognize Arabic characters.

The rest of paper is organized as follows: Section 2 presents some of the related work. Section 3 presents an overview of the DTW. The proposed Arabic OCR method is explained in Section 4. Section 5 illustrates results. Finally, conclusion and future work is presented in section 6.

2. RELATED WORK

Arabic OCR systems were not well studied as thoroughly as Latin, Chinese, or Japanese that have been paid more attention than both Arabic and non-Arabic-speaking researchers. This section lists some of several studies that were conducted on typed Arabic OCR.

In [18], authors proposed an Arabic OCR system that recognizes isolated typed characters, their proposed OCR system uses Zernike and Invariant

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

moments with Walsh transformation, their results reported 98% recognition rate.

In [19], authors proposed a method that is able to discriminate handwritten and typed Arabic texts using a bag of visual words model, after segmenting the Arabic text into blocks, their algorithm represents each block as a word vector, these word vectors include features that are recognized by a scale invariant feature transform (SIFT) algorithm [20], finally, their proposed algorithm decides whether the Arabic text block is handwritten, typed, or noise using a support vector machine classifier, their method shown a promising performance using a reliable evaluation approach evaluated on different datasets.

In [21], authors proposed a recognition technique for Arabic printed texts, their recognizer uses hidden Markov and bootstrap models, the usage of the two models optimizes the boundary of Markov model and enhances the performance of the proposed recognizer, their recognition technique decreases error rates for both word and character recognition to 13.3% and 14% respectively.

In [22], the authors investigated the usage of k-Nearest Neighbor (kNN) and Random Forest Tree (RFT) classifiers in Arabic OCR tasks, their OCR system performs a binarization process on the input images that contain the Arabic typed texts, extracts statistical features calculated on the shapes of the Arabic characters, and finally, evaluated the performance of the mentioned two classifiers, and, it is concluded that RFT classifier is much better than kNN by more than 11 % in recognition rate.

In [23], authors proposed an automatic recognition of fonts before starting OCR steps, in order to enhance the recognition accuracy of their Arabic OCR system, font recognition is achieved using SIFT descriptors [20] and a statistical feature selection algorithm is used with RFT classifier in the classification step of OCR systems, moreover, their text font recognition preprocessing step was helpful to achieve 99.8% to 100% character recognition accuracy when testing their proposed approach using a noise free Arabic OCR dataset of 30000 samples.

In [24], authors used a bidirectional long short-term memory (BLSTM) networks to implement a segmentation-free OCR system, BLSTM is a sort of recurrent neural networks that requires training by a pre-segmented data, and post-processing to transform outputs into label sequences. In their OCR system, a connectionist temporal classification method is used with BLSTM to label the unsegmented sequences, the proposed OCR system is evaluated using cursive Urdu and non-cursive English scripts and achieved 99.17% accuracy on non-cursive Latin scripts, 88.94% accuracy on Urdu cursive scripts without positive information, and 88.79% on Urdu cursive with position information, as a result, their approach requires more enhancements with cursive scripts, moreover, it needs more enhancements to work robustly with typed documents.

In [25], an automatic recognizer for printed Arabic texts is proposed, the recognizer uses linear and ellipse regression techniques, after gathering the possible forms of Arabic letters, their proposed recognizer generates a special code to represent each form for all the Arabic letters, the generated codes are used to identify characters. Moreover, the proposed recognizer is able distinguish fonts as it collects codes that are uniquely exist in the different fonts. Their proposed OCR approach is evaluated using 14000 different Arabic words with different fonts and achieved 86% recognition accuracy.

Arabic is among the most popular languages in the world. Hundreds of millions of people are speaking Arabic, however, because of its complexity, the recognition of printed and handwritten Arabic documents remains a challenge. Although, there are many published methods for that recognizes printed Arabic characters, there still needs to enhance recognition rate and speed for Arabic OCR systems.

3. DYNAMIC TIME WARPING ALGORITHM

DTW [8,10-14] measures similarity between two-time series and finds the optimal alignment between them by comparing a set of sample points of the two times sequences. The two sequences are usually warped in a non-linear fashion. Moreover, DTW may impose restrictions on the matching process to improve matching results and to reduce comparison complexity.

DTW may place restriction on the data sequences by sampling them at equidistant points in time which can be easily achieved by re-sampling, i.e. it may calculate the distance between the two-time series by resampling one of the series and by computing the distance sample by sample, however, this process may match samples that do not correspond well to each other.

It is known that DTW matching algorithm uses dynamic programming (DP) to align two time series, so that a distance metric is minimized. To formulate DTW, the following two time series S and T are defined in Equation 1 and Equation 2[26]:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$S=s_1,..., s_n \quad (1)$$

and

$$T=_{t1},\ldots,t_m \quad (2)$$

These two sequences S and T are arranged by DTW to form a n-by-m grid, where each grid point, denoted by (i,j), corresponds to an alignment between s_i and t_i, a warping path, W (see Equation 3) aligns the elements of the two sequences S and T, such that the distance between them is minimum.

$$W=w_1,\ldots,w_k \quad (3)$$

Noting that W is a series of grid points, where each w_k corresponds to a point denoted as $(i,j)_k$.

DTW requires a distance measure to formulate DP problems, the following are two candidates for a distance measure $\mathfrak{X}$ between two times series elements that can be used to implement DTW, one of them is the magnitude of the difference, Equation 4, and the other is the square of the difference between the two sequences S and T, Equation 5:

$$\mathfrak{X}(i,j)=|si-tj| \quad (4)$$

$$\mathfrak{X}(i,j)=(si-tj)^2 \quad (5)$$

To calculate the distance between S and T using the distance measure $\mathfrak{X}$, Equation 6 is defined:

$$DTW(S,T)=\min_w[\sum_{k=1}^{p} \mathfrak{X}(w_k)] \quad (6)$$

In DP, it is needed to define variables that describe legal state transitions, these variables include a stage variable, state variables, and decision variables. The stage variable is simply the time and it essential to enforce a monotonic order on events, the state variables are the individual grid points, and, the decision variables are constraints on acceptable paths between grid points to reduce the search space of DTW, and to minimize the number of warping paths.

It is known that calculating matrix $\mathfrak{X}$ is time consuming, thereby time and space complexities in the implementation of DTW is increased. Therefore, it is essential to set the following restrictions:

- ✓ Monotonicity:
  This restriction is ensuring the monotonically ordering of points with respect to stage variable, i.e. $i_{k-1}\leq i_k$ and $j_{k-1}\leq j_k$.
- ✓ Continuity:
  This restriction is ensuring the confining of steps in the grid to adjacent points, i.e. $i_k-i_{k-1}\leq 1$ and $j_k-j_{k-1}\leq 1$.
- ✓ Warping window:
  This restriction ensures falling of points in a given warping window, i.e. $|i_k-j_k|\leq w$, where w is the window width (w>0).
- ✓ Slope constrain:
  This constrain is to avoid extremely large movements of acceptable warping paths.
- ✓ Boundary conditions:
  This constrain is to further reduce the search space of DTW by constraining endpoints, $i_1=1$, $j_1=1$ and $i_k=n$, $j_k=m$.

DP is implemented using the following recurrence relation as shown in Equation 7:

$$\epsilon(i,j)=\mathfrak{X}(i,j) + \min[\mathfrak{X}(i-1,j), \mathfrak{X}(i-1,j-1), \mathfrak{X}(i,j-1)] \quad (7)$$

Where the cumulative distance, $\epsilon(i,j)$, is the sum of the distance between elements specified by a point and the minimum of the cumulative distances of their neighbors. DP uses this symmetric formulation without re-calculating partial path distances. Upon completion, DTW finds the best warping path by tracing the lowest cumulative distance.

4. THE PROPOSED ARABIC OCR USING DTW

It is known that each character in Arabic comes in different shapes depending on its location of the word (start, middle, or end), Table 1. Moreover, Arabic characters have different shapes, if they are isolated and some of them are classified into same groups that may have the same main stroke with minor changes (some of them may have different number of dots or different locations of dots with respect the main stroke). The main implementation steps of the proposed Arabic OCR system based on DTW algorithm are shown in Figure 1.

In addition to the challenges of Arabic characters, there are a number of factors that affect the Arabic OCR quality, such as the scanner quality, scan resolution, type of printed documents (laser printer or photocopied) and paper quality [15,16]. Therefore, in this work, segmentation of Arabic characters is merged with the recognition step to overcome the ordinary segmentation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

problems such as the overlapping of sub words, different sizes of written characters, and the existence of connecting strokes with different length between the written Arabic characters. Moreover, a number of pre-processing steps are used to make the Arabic OCRing more robust and to make it easier for the proposed OCR system in order to operate in accurate manner. The proposed algorithm steps, as shown in Figure 1:

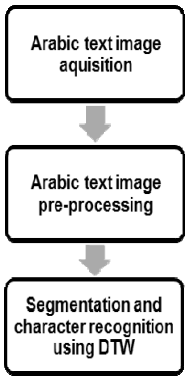

Figure 1. Implementation steps of the proposed Arabic OCR using DTW algorithm.

Step 1: Arabic text image acquisition:

In this work, a scanner is used to capture Arabic text images noting that the (color) images are scanned page by page using a 300 dpi scanner.

Step 2: Arabic text image pre-processing:

The scanned Arabic text pages are pre-processed [27], some of the scanned images suffer from slight rotation due to the scan process, and therefore, a rotation algorithm is applied to make them horizontally aligned. However, if the scanned images are color images, the OCR algorithm converts them to grey level images. In this work, the Arabic text image pre-processing incorporates the following detailed steps:

- ✓ Image binarization:
  The pre-processed grey images are binarized using Otsu's binarization algorithm [27]. Finally, binarized images are saved as bmp files.
- ✓ Noise reduction:
  The binarized images are subjected to noise reduction and enhancement processes. In image processing, there are many noise reduction approaches. In this paper, Median filter is applied on binarized Arabic text images to reduce noise. Median filters are one of the useful nonlinear digital filtering techniques, often used to remove noise such as reducing salt-and-pepper noise in digital images, it replaces a pixel by the median of all pixels in the neighborhood
- ✓ Thinning:
  Image thinning [27] is to convert a thick digital image into a thin digital image – a one-pixel width image, i.e. to obtain the skeleton form for each Arabic character that expresses the structural connectives of the main component of an Arabic character. The main advantage of a thinning process is to produce a more compact, convenient and condensed representation of a character image while preserving its topological properties. It is noted that thinning process is normally used with handwritten OCR systems, however, it is used with the proposed printed Arabic character OCR system to minimize the processing time for the recognition using DTW algorithm.

Step 3: Segmentation and character recognition using DTW:

The segmentation process is a basic OCR step, errors in segmenting characters produce errors in the recognition process in this work, the proposed segmentation-recognition process is based on pixel tracking is used to segment characters directly.

In this paper, segmentation [15,16] incorporates line and character segmentation steps that are embedded in character recognition using DTW. It is known that each Arabic text line consists of a group of words and each Arabic word consists of a group of characters that are adjacent to each other.

Arabic characters are sequentially divided into disjoint or overlapping cells for which size and location are not known in advance, this embedded character segmentation is inspired from profile-based representations of printed words presented in [28].

Our proposed recognition segmentation-recognition method captures the shape information of Arabic characters without any prior assumption on the font size or style used in the scanned text images, represents each character by a feature vector and matches it with other character profiles – templates. Finally, the algorithm forms Arabic words as a sequence of pattern cells that match profile-based feature vectors. The character shape matching is performed using DTW. DTW ensures great flexibility in the course of the matching process and handles local variations in the shapes of characters.

The proposed segmentation recognition using DTW incorporates the following detailed steps [17]:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- ✓ Line segmentation and cropping:
  The Arabic text binary images are segmented [15,16] into lines of text. Moreover, each segmented line is cropped to get rid of the white area around.
- ✓ *Cell height-width calculation*:
  The maximum vertical coordinate (MVC) and the base line coordinate (BLC) (see the upper and the lower red lines as shown in Figure 2) of cells are determined. The cell height is calculated by subtracting the BLC from the MVC.

Figure 2. Font size calculation [17].

In this work, the proposed OCR algorithm recursively increases the width of the cell, width increasing process continues until matching a candidate valid Arabic character, in this width calculation process, the proposed OCR algorithm sets the minimum character width (cell width) to 20 pixels.

If the cell width does not contain a valid character, it increases the cell width by 2 pixels at a time).
As Arabic characters are cursive, the cell width may contain a full character and a portion of another character, as shown in Figure 3.

To solve this problem, i.e. when containing a portion of the successor or predecessor character, the proposed Arabic OCR algorithm removes remnants of the other Arabic characters from the cell.

- ✓ *Candidate character normalization:*
  It is known that the sizes of Arabic characters may vary. As a result, size normalization is often used to scale candidate cells (characters) to a fixed size, i.e. the size of the cells must fit the size of the cell profiles (character templates). The OCR algorithm centers these candidate cells (characters) before executing the matching process using DTW. Cell (character) size normalization is one of the important pre-processing techniques that are commonly used in OCR systems.

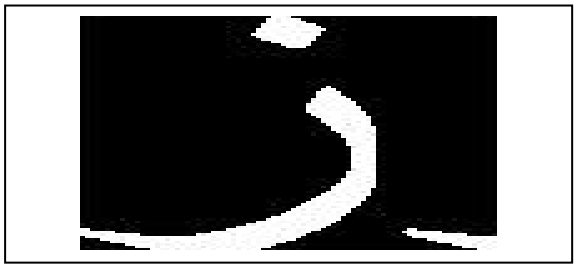

Figure 3. Character image with a remnant [17].

In this work, character (cell) sizes are normalized into 32×32 pixels and cropped.
The normalization process is performed as follows: Each segmented cell (character) is cropped, resized, and centered in a 32×32 pixel block before matching. Before recognition process, the normalized characters are middle position, initial position, end position, isolated character, or space between two words.

- ✓ *Character recognition using DTW*:
  Recognition is done by comparing the cell profiles of all character corresponding to the input cell (character) using DTW algorithm.
  These cell profiles - templates are prepared as follows: three randomly selected essays from Addustour[1], Jfranews2 and Al-Rai[3] newspapers are manually segmented. The segmentation process terminates after segmenting all the forms (initial, middle, final, and isolate) of the 28 Arabic characters from each of the three newspapers. The segmented characters templates – cell profiles that are resulted from the three newspapers are merged, however, highly similar of the duplicated character forms are discarded. These remaining cell (character) forms are normalized, cropped, resized and centered in a 32×32 pixel block. Notice that each Arabic character form has 1, 2 or 3 sample cell profiles - templates. These character forms are served as cell (character) profiles – templates in the online segmentation – recognition step. In each character line, the input – candidate cell (character) is searched from right to left. The search terminates whenever the classifier finds a meaningful character (or space). DTW finds the dissimilarity between two non-linear sequences potentially having different lengths, one of them is the candidate input cell (character) and the other is each of the Arabic character forms that are stored in previously collected character

[1] http://www.addustour.com/
[2] http://www.jfranews.com.jo/
[3] http://alrai.com/

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

cell profiles - templates. Finally, the algorithm returns the most similar character from the templates i.e. the matched character with the input character (An Arabic character is said to be OCRed correctly if it gives the shortest DTW distance with one of the characters in the cell (character) profiles - templates).

- ✓ *Arabic Character typing*:
  After character recognition, the proposed Arabic OCR using DTW types the recognized character in a text editor.

5. RESULTS

In order to evaluate the accuracy of the proposed Arabic OCR using DTW, we have used printed Arabic characters retrieved from Ad-Dustour newspaper, Alrai newspaper and Jafranews (newspapers in Jordan). These documents serve as our own database for printed Arabic text recognition of newspapers, and used to conduct experiments to evaluate the performance of the proposed OCR technique. These printed texts are scanned at 300 dots per inch using HP Scanjet 2400 scanner, Table 2 shows the statistics of the pages used in evaluating the performance of the proposed algorithm.

Table 2: Some useful statistics about the data used in evaluation of the proposed OCR system

Statistics parameter	Number
Pages	49
Words	50007
Characters (no spaces)	316360
Characters (with spaces)	370201
Paragraphs	6839
Lines	20794

Highly skewed text are discarded, the data in Table 2 was selected randomly from the mentioned Jordanian newspapers, so as to fairly evaluate the robustness of the proposed OCR approach, the whole dataset is subjected to noise removal and image enhancement processes, converted to binary images, cropped to get rid of the white area around the pages, normalized, thinned, segmented to lines, cropped to get rid of the white area around the lines, the font size of the candidate characters is determined, and compared through a recursive search to match one of the character forms of one of the predefined templates of Arabic characters using DTW algorithm. In this work, MATLAB (R2016.a/64-bit) is utilized to implement the proposed Arabic OCR using DTW algorithm on an HP Pavilion g6 machine (Intel(R) Core(TM) i5 CPU M480 @ 2.67GHz with 3.00 GB RAM) running a MS Win 10 64-bit operating system. It should be noted that the proposed Arabic OCR using DTW algorithm is implemented using a parallel MATLAB programming technique which is basically based on parallel MATLAB instructions such as *parfor*.

Figures 4 and 5 show the original text images, and the corresponding recognized Arabic text paragraphs, respectively, by using our proposed OCR module based on DTW.

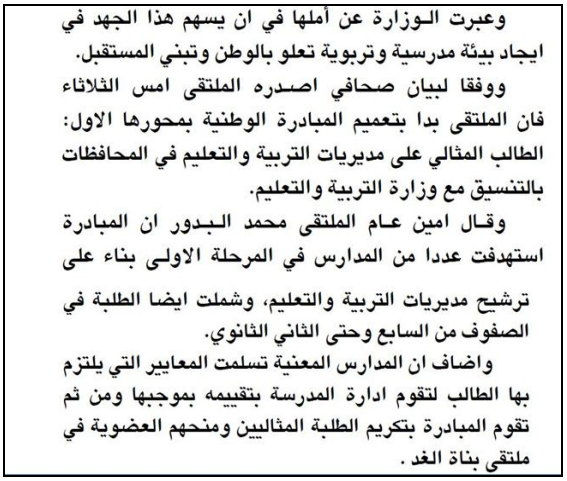
وعبرت الوزارة عن أملها في ان يسهم هذا الجهد في ايجاد بيئة مدرسية وتربوية تعلو بالوطن وتبني المستقبل.
ووفقا لبيان صحافي اصدره الملتقى امس الثلاثاء فان الملتقى بدا بتعميم المبادرة الوطنية بمحورها الاول: الطالب المثالي على مديريات التربية والتعليم في المحافظات بالتنسيق مع وزارة التربية والتعليم.
وقال امين عام الملتقى محمد البدور ان المبادرة استهدفت عددا من المدارس في المرحلة الاولى بناء على ترشيح مديريات التربية والتعليم، وشملت ايضا الطلبة في الصفوف من السابع وحتى الثاني الثانوي.
واضاف ان المدارس المعنية تسلمت المعايير التي يلتزم بها الطالب لتقوم ادارة المدرسة بتقييمه بموجبها ومن ثم تقوم المبادرة بتكريم الطلبة المثاليين ومنحهم العضوية في ملتقى بناة الغد .

Figure 4. The original text image – Example 1

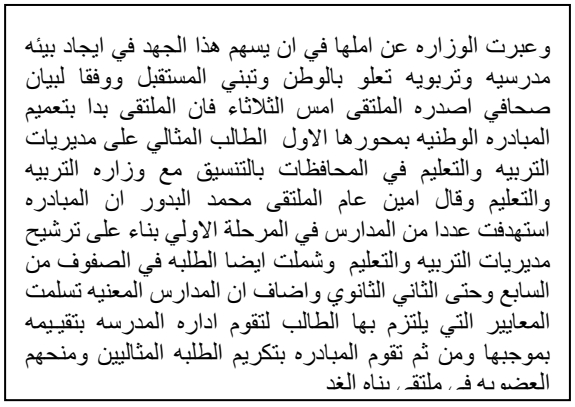
وعبرت الوزاره عن املها في ان يسهم هذا الجهد في ايجاد بيئه مدرسيه وتربويه تعلو بالوطن وتبني المستقبل ووفقا لبيان صحافي اصدره الملتقى امس الثلاثاء فان الملتقى بدا بتعميم المبادره الوطنيه بمحورها الاول الطالب المثالي على مديريات التربيه والتعليم في المحافظات بالتنسيق مع وزاره التربيه والتعليم وقال امين عام الملتقى محمد البدور ان المبادره استهدفت عددا من المدارس في المرحلة الاولي بناء على ترشيح مديريات التربيه والتعليم وشملت ايضا الطلبه في الصفوف من السابع وحتى الثاني الثانوي واضاف ان المدارس المعنيه تسلمت المعايير التي يلتزم بها الطالب لتقوم اداره المدرسه بتقييمه بموجبها ومن ثم تقوم المبادره بتكريم الطلبه المثاليين ومنحهم العضويه في ملتقى بناه الغد

Figure 5. An Arabic OCR using DTW output for the text image in Figure 4

In the same manner, Figures 6 and 7 show the original text images, and the corresponding recognized Arabic text paragraphs, respectively. Clearly, recognition process of our proposed technique is novel and robust.

To conduct a subjective comparison of text recognition, the proposed approach is compared with another printed Arabic text recognition system [17]. Authors in [17] implemented a correlation based OCR system. It is concluded that the proposed Arabic OCR algorithm using DTW works

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

much better. i.e. the average character recognition rate (CRR) of the proposed OCR algorithm using DTW is 98.0%, which is much better than the CRR of our previous OCR algorithm using correlation [17], where its best CRR is 96.2%. On the other hand, the worst CRR of the proposed OCR system using DTW is 96.2%.

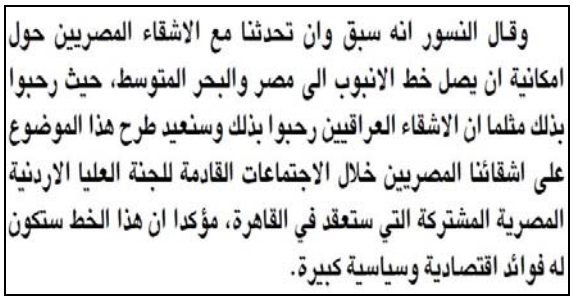
وقال النسور انه سبق وان تحدثنا مع الاشقاء المصريين حول امكانية ان يصل خط الانبوب الى مصر والبحر المتوسط، حيث رحبوا بذلك مثلما ان الاشقاء العراقيين رحبوا بذلك وسنعيد طرح هذا الموضوع على اشقائنا المصريين خلال الاجتماعات القادمة للجنة العليا الاردنية المصرية المشتركة التي ستعقد في القاهرة، مؤكدا ان هذا الخط ستكون له فوائد اقتصادية وسياسية كبيرة.

Figure 6. The original text image – Example 2

وقال النسور انه سبق وان تحدثنا مع الاشقاء المصريين حول امكانيه ان يصل خط الانبوب الى مصر والبحر المتوسط حيث رحبوا بذلك مثلما ان الاشقاء العراقيين رحبوا بذلك وسنعيد طرح هذا الموضوع على اشقائنا المصريين خلال الاجتماعات القادمه للجنه العليا الاردنيه المصريه المشتركه التي ستعقد في القاهره مؤكدا ان هذا الخط ستكون له فوائد اقتصاديه وسياسيه كبيره

Figure 7. An Arabic OCR using DTW output for the text image in Figure 6

Clearly, our proposed Arabic OCR using DTW works well for Arabic text images, however, some OCR problems are encountered in some of the Arabic characters such as Haa and Taa marbuta which are normalized to Haa. Same problem is encountered in Alif and its different shapes which are normalized to Alif without Hamza and in Yaa and Alif maksura which are normalized to Yaa.

6. CONCLUSION

The proposed Arabic OCR using DTW aims to transform the image of a scanned printed Arabic text into an editable electronic file. The proposed algorithm is implemented using parallel MATLAB code, it is based on the common OCR processes, it scans an Arabic document, enhances it by applying some of the image processing techniques, segments the Arabic image into lines of text, and then into characters. Rather than using the common segmentation practice, this work proposes a segmentation of Arabic texts into lines and characters. The proposed Arabic OCR overlaps the segmentation and the recognition processes, in order to avoid the challenges of segmenting the highly cursive Arabic texts to isolated characters.

Although significantly high recognition results can be accomplished for Arabic typed text especially for a single font type, many challenges must to be considered by this paper in the future: mixed-font Arabic newspaper text recognition shall be addressed [29,30], a font identification is to be considered in preprocessing phase, a skew correction technique [31] is to be implemented, and finally the examination of post-processing methods such as spelling correction [32] that may improve character recognition rates is also can be considered.

ACKNOWLEDGEMENTS

Authors like to thank MEng. Nawal Al-Zaben for her valuable comments and advice.

REFRENCES

[1] Adnan Amin, "Off line Arabic Character Recognition-A Survey", *Proceedings of the Fourth International Conference on Document Analysis and Recognition,* Vol. 2, 1997, pp. 596-599.

[2] Horst Bunke, Patrick Shen-pei Wang, (Editors), "Handbook of Character Recognition and Document Image Analysis", *World Scientific*, 1997.

[3] Mohammad Khorsheed, "Off-Line Arabic Character Recognition – A Review", *Pattern Analysis & Applications*, Vol. 5, 2002, pp. 31-45.

[4] Ahmed Lawgali, "A Survey on Arabic Character Recognition", *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 8, No. 2, 2015, pp. 401-426.

[5] Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, and Soumya Ghosh, "Optical Character Recognition Systems for Different Languages with Soft Computing", *Studies in Fuzziness and Soft Computing 352*, Springer International Publishing, 2017.

[6] Abdelwadood Mesleh, "Support Vector Machine Text Classifier for Arabic Articles", *VDM Verlag Dr. Müller*, 2010.

[7] Oleg Redkin, Olga Bernikova, "On the Optical Character Recognition and Machine Translation Technology in Arabic", *Proceedings of the 2011 International Conference on Artificial Intelligence*, Las Vegas, USA, Jul. 18-21, 2011, pp. 861-867.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[8] Joseph B. Kruskal, Mark Liberman, "The Symmetric Time-Warping Problem: From Continuous to Discrete. In Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison", Editted by David Sankoff, Joseph B. Kruskal, *Addison-Wesley Publishing Co., Reading, Massachusetts*, 1983, pp. 125-161.

[9] Maher Khemakhem, "Arabic Type Written Character Recognition Using Dynamic Comparison", *Proceedings of 1st Computer Conference*, Kuwait, March 1989, pp. 109-118.

[10] Hamdi Hassen, Maher Khemakhem, "Large Distributed Arabic Handwriting Recognition System based on the Combination of FastDTW Algorithm and Map-reduce Programming Model via Cloud Computing Technologies", *AASRI Procedia*, Vol. 5, 2013, pp. 156-163.

[11] Stan Salvador, Philip Chan, "FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space", *KDD Workshop on Mining Temporal and Sequential Data*, 2004, pp. 70-80.

[12] Radu Prodan, Simn Ostermann, "A Survey and Taxonomy of Infrastructure as a Service and Web Hosting Cloud Providers", *Proceedings of the International Conference on Grid Computing*, Banff, AB, Canada, Oct. 13-15, 2009, pp. 1-10.

[13] Zied Trifa, Mohammed Labidi, and Maher Khemakhem, "Dynamic Time Warping Algorithm with Distributed Systems", *World of Computer Science and Information Technology Journal*, Vol. 1, No. 4, 2011, pp. 132-137.

[14] Yu-Liang Hsu, Cheng-Ling Chu, Yi-Ju Tsai, and Jeen-Shing Wang, "An Inertial Pen With Dynamic Time Warping Recognizer for Handwriting and Gesture Recognition", *IEEE Sensors Journal*, Vol. 15, Issue 1, 2015, pp. 154-163.

[15] Ahmed M. Zeki, Mohamad S. Zakaria, and Choong-Yeun Liong, "Segmentation of Arabic Characters: A Comprehensive Survey", *International Journal of Technology Diffusion*, Vol. 2, Issue 4, 2011, pp. 48-82.

[16] George Kour, "Real time segmentation and recognition of on line handwritten Arabic scripts", *Master thesis*, Tel Aviv University, 2014.

[17] Abdelwadood Mesleh, Ahmed Sharadqh, Jamil Al-Azzeh, Mazen Abu-Zaher, Nawal Al-Zabin, Tasneem Jaber, Aroob Odeh, and Myssa'a Hasn, "An Optical Character Recognition", *Contemporary Engineering Sciences*, Vol. 5, No. 11, 2012, pp. 521-529.

[18] Mustapha Oujaoura, Rachid El Ayachi, Mohamed Fakir, Belaid Bouikhalene, and Brahim Minaoui, "Zernike moments and neural networks for recognition of isolated Arabic characters", *International Journal of Computer Engineering Science*, Vol. 2, Issue 3, 2012, pp. 17-25.

[19] Konstantinos Zagoris, Ioannis Pratikakis, Apostolos Antonacopoulos, Basilis Gatos, and Nikos Papamarkos, "Distinction between handwritten and machine-printed text based on the bag of visual words model", *Pattern Recognition*, Vol. 47, Issue 3, 2014, pp. 1051-1062.

[20] Andrey Rosenberg, "Using SIFT Descriptors for OCR of Printed Arabic", *Master thesis*, Tel Aviv University, 2012.

[21] Zhiwei Jiang, Xiaoqing Ding, Liangrui Peng, and Changsong Liu, "Modified Bootstrap Approach with State Number Optimization for Hidden Markov Model Estimation in Small-Size Printed Arabic Text Line Recognition", *Proceedings of the 10th International Conference of Machine Learning and Data Mining in Pattern Recognition*, St. Petersburg, Russia, Jul. 21-24, 2014, pp. 437-441.

[22] Marwa Rashad, Noura A. Semary, "Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers", *Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications*, Cairo, Egypt, Nov. 28-30, 2014, pp. 11-17.

[23] Mohamed Dahi, Noura A. Semary, and Mohiy M. Hadhoud, "Primitive Printed Arabic Optical Character Recognition using Statistical Features", *Proceedings of the 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS'15)*, Cairo, Egypt, Dec. 12-14, 2015. pp. 81-86.

[24] Saad Bin Ahmed, Saeeda Naz, Muhammad Imran Razzak, Shiekh Faisal Rashid, Muhammad Zeeshan Afzal, and Thomas M. Breuel, "Evaluation of cursive and non-cursive scripts using recurrent neural networks", *Neural Computing and Applications*, Vol. 27, No. 3, 2016, pp. 603-613.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[25] Ashraf A. Shahin, "Printed Arabic Text Recognition using Linear and Nonlinear Regression", *International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 1, 2017, pp. 227-235.

[26] Donald J. Berndt, James Clifford, "Using dynamic time warping to find patterns in time series", Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Seattle, USA, Jul. 31- Aug. 01, 1994, pp.359-370.

[27] Rafael C. Gonzalez, Richard E. Woods, "Digital image processing", *Pearson Education*, 2008.

[28] Million Meshesha, C. V. Jawahar, "Matching word images for content-based retrieval from printed document images", *International Journal on Document Analysis and Recognition*, Vol. 11, No. 1, 2008, pp. 29-38.

[29] Rohit Prasad, Shirin Saleem, Matin Kamali, Ralf Meermeier, and Prem Natarajan, "Improvements in hidden Markov model based Arabic OCR", *Proceedings of 19th International Conference on Pattern Recognition*, Tampa, FL, USA, Dec. 8-11, 2008, pp. 1-4.

[30] Kamel Ait-Mohand, Thierry Paquet, and Nicolas Ragotm, "Combining structure and parameter adaptation of HMMs for printed text recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, pp. 1716-1732.

[31] Irfan Ahmad, "A Technique for Skew Detection of Printed Arabic Documents", *Proceedings of the 10th International Conference Computer Graphics, Imaging and Visualization*, Macau, China, Aug. 6-8, 2013, pp. 62–67.

[32] Mariam Muhammad, Tarek ELGhazaly, Mostafa Ezzat, and Mervat Gheith, "A Spell Correction Model for OCR Errors for Arabic Text", *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*, Cairo, Egypt, Oct. 24-26, 2016, pp.124-136.