Skew Detection/Correction and Local Minima/Maxima Techniques for Extracting a New Arabic Benchmark Database

Husam Ahmed Al Hamad Department of Information Technology Qassim University Qassim, Saudi Arabia

Abstract—We propose a set of techniques for extracting a new standard benchmark database for Arabic handwritten scripts. Thresholding, filtering, and skew detection/correction techniques are developed as a pre-processing step of the database. Local minima and maxima using horizontal and vertical histogram are implemented for extracting the script elements of the database. Elements of the database contain pages, paragraphs, lines, and characters. The database divides into two major parts. The first part represents the original elements without modifications; the second part represents the elements after applying the proposed techniques. The final database has collected, extracted, validated, and saved. All techniques are tested for extracting and validating the elements. In this respect, ACDAR proposes a first issue of the Arabic benchmark databases. In addition, the paper confirms establishment a specialized research-oriented center refers to learning, teaching, and collaboration activities. This center is called "Arabic Center for Document Analysis and Recognition (ACDAR)" which is similar to other centers developed for other languages such as English.

Keywords—ACDAR; Arabic benchmark database; Arabic scripts; document analysis; handwriting recognition; skew detection and correction

I. INTRODUCTION

Arabic language is spoken by hundreds of millions of people around the world. It profoundly influenced many cultures, including the Western culture, for many centuries. Although it is one of the most important languages in the world throughout its long history, it still lags behind many other languages as far as information technology resources and applications are concerned. As a result, the so-called "digital gap" is greater for Arabic language than other languages such as English, for instance.

Automatic recognition of handwritten words remains a challenging task even though the latest improvements of recognition techniques and systems are very promising. The term handwriting refers to some artificial graphical marks containing a message in a given human language [1]. The concept of handwriting has always existed, for the purpose of expanding people's memory and facilitating communication together [2] and much of the human culture may be attributed to the advent of handwriting. Because of the fact that only humans can perfectly understand and recognize the handwritings of others, one computationally challenging task

resides in the attempt to imitate the human ability to read and recognize handwriting [3]. Consequently, automatic recognition of handwritten words remains a difficult task even though the latest improvements of recognition techniques and systems seem to be promising. For the purpose of automating Arabic scripts processing, numerous contributions have made in the area of handwritten script segmentation and recognition [4]. However, no outstanding results were reached so far, as OCR Arabic processing is still facing serious issues. One of the reasons is that Arabic language is considerably harder than Latin counterpart [5]. Therefore, in the area of automatic recognition of Arabic handwriting, many works have still to be done. One of the most important requirements for the development and comparison of recognition systems is a large database together with ground truth information. Compared to Latin scripts where handwritten words and numbers have publicly available for a long time (e.g. CEDAR [6], NIST¹) the situation for Arabic is quite different. Others implement large databases that are not available to the public [7], or unreliable databases that concern only one Arab country (e.g. IFN/ENIT [8]).

Although many research efforts have done, so far in the recognition of handwritten Arabic script [4] until now they have not reached satisfactory results for the following reasons [2].

- Arabic words are overlapped and written always cursively, i.e., more than one character can be written connected to each other.
- Arabic writing uses many types of external objects, such as 'dots', 'Hamza', 'Madd', and diacritic objects, these external objects make the task of line separation and segmentation scripts more difficult.
- An Arabic character can have more than one shape according to its position in the word, i.e., initial, middle, final, or as a standalone character.
- Arabic writing uses many ligatures, especially in handwritten text.
- Other characters have very similar contours and are difficult to segment and to recognize especially when non-characters and external objects are present in the

¹ NIST database, http://www.nist.gov/srd/

scanned image.

II. HISTORICAL BACKGROUND

Earlier surveys discussed recognition and segmentation of both handwriting and machine-print, with much emphasis on machine-print. Unfortunately, only a small and unreliable database is available for Arabic Language today. However, in 1980, Nouh et al. suggested a standard Arabic character set to facilitate computer processing [9]. Standard and reliable databases were developed many years ago for the recognition of handwriting in Latin scripts. Among these databases, the CEDAR database (Center of Excellence for Document Analysis and Recognition) was released in 1993 [6]. It contains images of approximately 50,000 alphanumeric characters, 5,000 city names, 5,000 state names, and 10,000 ZIP codes. Each image was scanned from mail in a working post office at 300 pixels per inch in 8-bit grayscale on a high quality flatbed digitizer. The data were unconstrained for the style and technique of preparation. writer, These characteristics help overcome the limitations of earlier databases that contained only isolated characters or were prepared in a laboratory setting under prescribed circumstances. In addition, the database is divided into explicit training and testing sets to facilitate the sharing of results among researchers as well as performance comparisons.

In 1999 Al ISRA Arabic database [10] collected from 500 students, it contains words, digits, signatures, which is has limitation because it does not contain paragraphs. Another database lunched in 2002 is IFN/ENIT [4, 11], it was developed at the Institute of Communications Technology (IFN) at Technical University Braunschweig in Germany and the Ecole Nationale d'Ingenieurs de Tunis (ENIT) in Tunisia. It consists of 26,459 images of the 937 cities names and towns in Tunisia, written by 411 different persons filled forms with about 26400 names containing more than 21,0000 characters. For each name some information are coded such as the sequence of character shapes, some style information, and the baseline are coded. It is used for recognition of data entry, mail sorting, and other recognition tasks. The images are partitioned into four sets so that researchers can use and discuss training and testing data in this context. The database has certainly many advantages and some of its drawbacks is the fact that it is written for Tunisia only and therefore contains only Tunisian cities and names and does not cover other Arab countries. It also lacks reliable training and testing sets. As a result, it is not widespread among researchers.

One of the efforts that addressed the handwriting recognition problem is in writing personal checks. One such system, developed a decade ago, is AHDB (Arabic Handwritten DataBase), a database containing samples from 100 different writers, including words used for numbers [12]. In 2003 AI-Ohali et al., developed CENPARMI images databases from 3,000 checks and implemented at the Center for Pattern Recognition and Machine Intelligence provided by a banking corporation [13].

These databases contain numeric amounts written in words, sub-words, Indian digits, and numeric amounts written with Indian digits. Notably, Indian digits are the numeric digits normally used in Arabic writing, as opposed to "Arabic numerals" ordinarily used in Latin script. The Indian digits database contains 15,175 samples, the legal and courtesy databases 2,499 samples, and the sub-words database contains 29,498 samples.

In 2009 ADAB database (Arabic DAtaBase) with Arabic online handwritten words has used by Haikal, *et al* [14] at the first time, the database was developed for Arabic online handwritten scripts in a cooperation between the Institute for Communications Technology (IfN) and the Ecole Nationale d'Ing'enieurs de Sfax (ENIS). The database written by more than 130 persons, it consists of 15158 Arabic handwritten words, 937 Tunisian town/village names. The database contains in additional special tools for the collection of the data and verification of the ground truth. These tools give the possibilities to record the online written data, to save some writer information, to select the lexicon for the collection, and re-write and correct wrong written text.

Although the recognition accuracy for separated handwritten numerals and characters has improved significantly in recent years, the final frontier remains the accurate recognition of handwritten Arabic scripts. The pursuit of more accurate recognition rates continues to encourage researchers in the field. It must also be mentioned that along with the challenging nature of the handwritten word recognition problem, immense potential lies in the commercial sector to make these systems available. So, an important bulk of work is required to undertake a serious research and meet its multiple challenges.

III. BENCHMARK DATABASE

One of the most important components in ACDAR center is the benchmark database; often recognition algorithms have tested using one type of database, especially in the case of offline handwriting recognition. ACDAR is concerned with both off-line and online handwriting recognition. It proposes a common benchmark database of Arabic handwritten scripts, which is essential for research on handwritten Arabic word recognition. The first issue of this database has hosted in the center.

ACDAR began to work with the off-line Arabic handwriting recognition, which is may divided into segmentation-based and holistic ones. In general, the former approach uses a strategy based on the recognition of individual characters or patterns whereas non-segmentation based deals with the recognition of the word image as a whole [15]. In the online case, the handwriting has captured and stored in digital form via different means. Usually, a special pen has used in conjunction with an electronic surface. As the pen moves across the surface or paper, the two-dimensional coordinates of successive points have represented as a function of time and have stored in order [1]. It is generally, information is not easy to recover from handwritten words written on a non-digital medium such as accepted that the online technique of recognizing handwriting has so far achieved better results than off-line. This may be attributed to the fact that more information may be captured in the online case such as the direction, speed and the order of strokes of the handwriting. At the end, ACDAR's database will be made freely available to researchers.

A. Data collection

ACDAR started recognizing the paragraphs, lines, words, and characters. The handwriting papers have wrote by 113 distinct writers and scanned in a RGB-scale. The writers are variant in age, education, background, genders, and countries; Figure 1 shows a snapshot form that contains the instruction and personal details of the writers, Table I shows the statistical data that collected from all writers. Figure 2 illustrates the comparisons of database contains.

As a result, two paragraphs contain all shapes of Arabic characters have wrote by those writers. Figure 3 shows the original two paragraphs have requested to write by the persons, Figure 4 displays in blue the position of the different characters shapes, the second paragraph has required for collecting more samples.

www.acdar.org	ARABIC SOLUTIONS
	لأخ العزيز، الأخت العزيزة
	لسلام عليكم ورحمة الله وبركاته ، ، ،
بية، نرجو التكرم بكتابة الفقرتين	لغايات استكمال أحد البحوث العلمية بموضوع التعرف على الخطوط المكتوبة باللغة العر
حيت تكون كل ففرة على صفحه ختياري) في المكان المحدد أدناه،	الموضـحتين ادناه بخط واضح ويتسكل متاني ويخط مستفي ع على الورقتين المرفقتين بـ منفصلة، مع ضرورة تحديد العمر والجنس وبلد النشأة، والشهادة العلمية وكتابة الاسم (ا
	علماً بأن كافَّة المعلومات ستكون سرية ولغايات البحث فقط.
	واقبل/ اقبلي فانق الشكر ، ، ،
	واقبل/ اقبلي فائق الشكر ، ، ، ريق البحث
	واقبل/ اقبلي فائق الشكر ، ، ، ريق البحث حلومات الكاتب:
ے ذکر <u>ا</u> آنٹی	واقبل/ اقبلي فاتق الشكر ، ، ، ريق البحث حلومات الكاتب: الحم (مدين: الجنسي:
🗆 نکر 🗆 آنٹی	واقبل/ اقبلي فاتق الشكر ، ، ، ريق البحث مطومات الكاتب: لاسم ومدين: محســــر: [الل من ١٨ [19 ــ ٢٩ [٢١- ٥٥] اكثر من ٥٥ يبد النشأة.

Fig. 1. Form of the personal details of each writers



TABLE I. STATISTICAL DATA OF 113 WRITERS

Fig. 2. Comparisons of the statistical, a) education, b) age, c) gender, and d) country

c) gender

Lines, words and characters have also extracted and saved in the database; verification phase has also investigated before the final adoption of the samples for quality purposes. In summary, each writer has wrote 358 words (first paragraph contains 162 words, second paragraph contains 196 words), both of them are 1,916 characters, on average each writer has wrote 30 lines. In total, all writers have wrote 226 paragraphs, 3,390 lines, 40,454 words, and 216,508 characters. Number of 57 writers are identified what they wrote as a training set, also 56 are writers identified what they wrote as a testing set. Table II illustrates the numbers of images have collected before extraction and validation processes.

d) country

الفقرة الأولى: الفقرة ذاتيلي: وقال الذلقي بابم مديرة البر امج التطبيبية في مايكروسوفت جنوب الخليج وبمنظور علمي فإن شمول ووضوح أساس وقال الذلقي بابم مديرة البر امج التطبيبية في مايكروسوفت جنوب الخليج وبمنظور علمي فإن شمول ووضوح أساس البة التربيب وطرق جديدة لمصلق وريط مهار اتها في محل تقتيات المعلومات والاتصالات ومراعاة ظروف عملهم و ضر معار فهم من خلالها وابتكار فرص أكثر ارتباط بواسطة فرش عملية تنص عليها الاتصالات ومراعاة ظروف عملهم و ضر معار فهم من خلالها وابتكار فرص أكثر ارتباط بواسطة قرش عملية تنص عليها الاتصالات مراعاة ظروف عملهم و ضر معار فهم من خلالها المتكار فرص أكثر ارتباط بواسطة أورش عملية المعلم عليها الاتصالات في معارفي من شائعة أن ترش م المتواصل الحواء المنظفة ودح مميرة التعليم فيها بتوفير أفضل العربية إلى هذه المبادرة لاتها تكمن الكثرام الميكروسوف المواصل الحواء المنظفة ودح مميرة التعليم فيا بتوفير أفضل العربية الي هذه المبادرة لاتها تكر من شائعة أن تلاري والمروفة التقنية اللازمة، و تخليلين المعالم من من دخول المعرف العمل التعليم بهاينا المعالم المواصلة و منهم قرورة المعام الميكن المقالم من المعان المواحين على مرض وسائع المعالم المبيرة التعليم والميز و المعالين المعالم الميان المبيرة المعالم الميزة المراح علين المير المعام من دخول المعرف الميان المبيرة المعام المعام المبيرة المبيرة المعام المعام المبيرين المواحية المعام معار المعام المبيرة المعام معار المعام المبيرة المعام المبيرة المعام المبيرة الم والمروفة التفار ومنزة أورم الماليم عليم مع من دخول المعرف العمل المالية، إلى خلك نقول أن معتلي فرور العام البيث أن أحدث تغيير في المتام الثاني وابعاد دماع الشابي من الهامش الذي يمكن أن ينتج عنه متيني فرور العام البيث أن أحدث تغيير في المتام الثاني وابعاد دماع الشبياب من الهامش الذي يمكن أن ينتج عنه الخطو 14 41

لفقرة الثانية:

الفقرة الشقية: وتمتكل مبادرة "شركاه في التعليم" التي أطلقتها مايكروسوفت عالميا في أولون لذليلا واضحا على إنراك مايكروسوفت لأهمية التعليم الثقني في المدارس، وسوف توفر المبادرة التي تعليق مطلبا بمنظور عالمي استثمارات بر معينة هامة ومنحا يتفيدة تصل فيتها إلى مائتين وخمسون مليون دو لار على مدى السنوات الشمس المقبلة لتدريب الطلاب والمعلمين على مهارات تقنية المعلومات والاتصالات وتوفير الدما الثقني اللازم لهم. وتوفر المبادرة والتي منافية مطلبا بمنظور عالمي المي والمعلمين أض مايكر وسوفت مهارات تقنية المعلومات والاتصالات وتوفير الدما الثقني اللازم لهم. وتوفر المبادرة باللطلاب والمعلمين أيضا مجموعة من مكومة ومؤسسات ومنظموات خالفامية الخاصة بسطح المكتب وبأسفار مخفضة وريتم تطلبي والمعلمين أيضا مجموعة من محكومة ومؤسسات ومنظموات خاصة بسلح المكتب ومنعارت المجليمة المعلية لذلي بلد من جهنة قال معد الطوف محمد مدير عام مايكروسوفت التعليمية الخاصة بسلح المكتب وماحة المجلمة المعلية لذل بلد. من جهنة قال معد اللطوف الحكومات والمطومات في المعليمة الخاصة بموحل علم مدين محمد مدير عام مايكروسوفت التعليمية، وعليم ملورت الشركة معادرة "الثركاء في التعليم" المولية لذل بلد. من جهنة قال معد اللطوف الحكومات والمؤدست المطومات في المعليمة، وعليم ملورت الشركة معادرة "اشركاء في التعليم" بدعم وتليد كبرين من الحكومات والمؤدسات الثقافية، مركزة على أمعية المورت الشركة معادرة "شركاء في التعليم" المعلية التي را تعليم الحكومات والمؤدسات المعلمين على تطوير معان المورة الطري لتاتريس. وأضل إلى أن مبادرة الشركاء في القعليم التي ولم إلى مساعدة الأمل النجام الألمليون على قطور مع المدى الطوبل تلتي في مساب الرزية والمولية التريس المرابة المرابية المرابية والمؤدة والهادفة الطلاب على نهل المام والملمان على تطوير معوان المولي الترس. وأمل إلى أن مايكر وسوف على المؤلم والفر الم الم الم مالم رميزان والمؤدسات الشركات على المدى الطوبل تلتي في طساب الرزية المراب المرابة والمولية المولي وليون المرابة والم فرابي المزية وقر ها ملكر معام الموات والشركات على المين المائمة بالاعتماد على الأدوات والتقرب التري في الم مايكر وسوفت

Fig. 3. Two paragraphs cover all shapes of Arabic characters

الفقرة الأولى: تغطى كافة حالات أحرف اللغة العربية... وقال الذاطق باسم عنين البراحية التطيفية في مايكر وسوفت جنوب الذليج وبمنظور علمي فان شمول ووضوح أساس الانفاقية الذي أبر منه مع زار ان التطيم في المنطقة سوف تتيح لمنات الاف الطلاب في العالم العربي فوصة أكبر لضبط الية التدريب وطرق جنيدة لصقل وربط مهار اتيم في مجال تقنيات المعلومات والاتصالات ومراحاة ظروف عملهم وغمر معارفهم من خلالها، وابتكار فوص أكثر ارتباط بواسطة ورش عملية تنص عليها الاتفاقية بما لا يقل عن مانتي ورشة، معلوفهم من كدلها، وابتلال ونص انقرا (تفاشع واسطه ورش عميه دنص عميه ادعمج بما يو يع محس ورسم. ونحن سعداء بتنسمام شركة أمواج في الحيد من الدول العربية إلى هذه المبادر لائيا تحكس القرام مكروسوفت المتواصل اتجاه المنطقة ودعم مسرد التعليم فيها بتوفير أفضل أنواع التفتيت والحلول التعليمية للتي من شانيه أن تثري تجرية التعليم وتعزز قدرة الطلاب على التعلم بما يمكنهم من دخول المعترك للعلي متسلحين بالمبادئ الأساسية للعلم والمعرفة الثقتية اللازمة، وتغنيض نسبة أمية التكنولوجيا لتثبيت عمل مرض وسلخ المفاهم السلبية، إلى نثلك نقول أن منها ثورة العلم ما لبث أن أحدث تغيير في المتاح التقالي الناشئ وإبعاد دماغ الشباب من الهامش الذي يمكن ان ينتج عنه

Fig. 4. Position of all Arabic characters shapes covered by only the first paragraph

TABLE II. NUMBERS OF COLLECTED IMAGES BEFORE EXTRACTION AND VALIDATION PROCESSES

Details / writers	Each writer	Training set by 57 writers	Testing set by 56 writers	Total (training and testing sets) 113 writers			
Paragraphs	2	114	112	226			
Lines	30 as average	1,710	1,680	3,390			
Words	358	20,406	20,048	40,454			
Characters	1.916	109.212	107.296	216.508			

The images have scanned in 200, 300 dpi resolution in RGB-scale images [2, 16, 17, 18]. Two version of the scanned documents have saved in the database, before and after preprocessing. Paragraphs, lines, words, and characters have extracted and saved as well.

The key steps of the techniques that were developed in this research is shown in Figure 5, the Figure shows briefly how we extracted all paragraph, lines, words, and characters. The first step is scanning the original documents with 200 and 300 dpi in RGB-scale, next step is preprocessing the scanned images, skew detection / correction, thresholding, and remove the noise using filtering are investigated in this stage. Next, start extracting process of the database; this stage includes developing a set of techniques to get the best extraction results of lines, words and characters. Finally, the last step is validating step, all extracted elements underwent to the evaluation process, if the element successfully passed this stage, then it will save into the benchmark database, otherwise it will discard. As mentioned before, this research aims to

build the first issue of ACDAR database and test the proposed algorithms have developed in this research.

B. Pre-processing

Many techniques have developed to perform further processing to allow superior recognition. Thresholding and filtering which they aim to eliminate and remove any noise or any small ascenders. Skew detection and correction technique that aims to adjust slopes of the paragraphs and lines. Next sub-sections are explain in details the parts of the preprocessing.

1) Thresholding and Filtering

The first step of preprocessing is thresholding (binary format); it uses as prior to further processing. Thresholding involves the conversion of a grey-scale image (0-255) into a binary image (0-1). This format will be easier to manipulate an image without levels of color in some researches, in additional the processing will be faster, less computationally expensive and will allow for more compact storage. The goal of using the thersholding is to determine the segmentation points of the lines, words, and the characters. Determine the segmentation points from the grey-scale image will be easier than color the image; the same points have extracted were applied on the RGB-scale. There are of course many of the defects such as loosing features from image. However, since the goal of this stage is only to determine the segmentation points, the effect will be the lowest grades possible. rgb2gray function was used to converts RGB images into grayscale by eliminating the hue and saturation information while retaining the illumination. The definition rgb2gray is shown in the following equations. The im2bw function was also used to convert this grayscale image to binary format (matrix). The output binary image BW has values of 0 as a foreground pixel (black) for all pixels in the input image and 1 as a background pixel (white) for all other pixels. All images were converted using the previous technique so that only binary images remained and could be used for further processing.

$$g = w_r I_r + w_g I_g + w_b I_b \tag{1}$$

s.t.
$$w_r + w_g + w_b = 1$$
, (2)

$$w_r \ge 0, w_a \ge 0, w_b \ge 0, \tag{3}$$

g is a constraint linear combination of R, G where, and B channels of input color image I,

- I_r , I_g , and I_b are the inputs,
- w_r, w_q , and w_b : weights sum to 1, and they

are non-negative numbers.

Next, elimination of the elements noise; the goal of this technique is to remove the noise as well as small foreground objects that were not part of the writing. Once the component of word image as matrix were identified, it was possible to perform various useful operations. *imfilter* function has used, it performs multidimensional filtering according to the specified options like *fspecial* function to create 2-D special filters that used 'disk' function to returns a circular averaging filter 'pillbox' within the square matrix of side (2 * radius + 1). Gaussians function [19] is applied, it was used at the lowest degree possible in order to not lose the features of the scripts as much as possible. The function aims to make the word image more smoothly, and to eliminate any small ascenders or

descanters noise between the lines. The following equation shows the one per direction using Gaussians function.



Fig. 5. Steps of extracting and validating the benchmark database

$$g(x,y) = \frac{1}{2\pi\sigma^2} \quad . \quad e^{-\frac{x^2 + y^2}{2\sigma^2}} \tag{4}$$

 $where, \sigma$ is the standard deviation of the Gaussian distribution,

x is the distance from the origin in the horizontal axis,

y is the distance from the origin in the vertical axis.

2) Skew Detection and Correction

Before extracting the lines, words, and characters from the documents images, skew of the paragraphs should be detect and then correct, the technique uses projections of an image matrix along specified directions, Hough transform [20] algorithm is applied to detect and correct the slopes. Hough Transform is the linear transform for detecting straight lines, the straight line is described as y = mx + b where the parameter *m* is the slope of the line, and *b* is the intercept (*y*-

intercept). Before start to apply the Hough transform algorithm, the document image should be prepared. So, a set of steps were used, at the beginning Threshold of the image to binary was applied. Next, in order to obtain a clear base line for all lines in the page, the punctuation marks (dots) and small stroke have removed from the image. Then, dilate image is also applied to close the internal gaps between the characters and words as well. Closing operation was performed upon the horizontal line element and merging the words of the lines. The text lines now look likes rectangles, to apply the Hough transform one-step is remained, this step is thinning the image includes all horizontal rectangles. To find the skew of the image, the mean and standard deviation of slopes were calculated, any bad data conceders far away from the standard deviation was removed, then the average of the good slopes was calculated, therefore the skew can be calculated by using the angle of the slope. The skew correction has applied using the negative of this angle. The below equations shows the Hough transform technique, the line

equation can be written as shown in equation (5), rearranged the equation shown in equation (6), and equation (7) shows formula of an point on the image with coordinates. Figure 6 shows the steps of skew detection and correction have developed in this research; Figure 7 shows samples of skew detection and correction for one paragraph and one line.

$$y = \left(-\frac{\cos\theta}{\sin\theta}\right)x + \left(\frac{r}{\sin\theta}\right) \tag{5}$$

$$r = x\cos\theta + y\sin\theta \tag{6}$$

$$r(\theta) = x_0 \cos \theta + y_0 \sin \theta \tag{7}$$

where, r distance between the line and the origin, is determined by θ ,

 $\boldsymbol{\theta}$ is the angle of the vector from the origin to this closest point.



Fig. 6. Steps of skew detection and correction

C. Extracting the database

Local minima and maxima of horizontal and vertical histograms have used for determining the segmentation points SPs for extracting the lines, words and characters. The concept of using the horizontal histogram is for extracting the line image. Horizontal histogram is formed by counting the total numbers of foreground pixels (black color) for each row from left to right in the paragraph image; the segmentation points have located based on the white color (background pixel) or the distance between two successive local maxima and one local minima with almost no foreground pixels. Figure 8 illustrates technique of extracting the lines images based on horizontal histogram.



Fig. 7. Sample of skew detection and correction, a) a paragraph b) a line

Using the same technique, but now by applying the vertical histogram to extract the word images. Vertical histogram has formed by counting the total numbers of foreground pixels (black color) for each column from top to bottom in the line image. The segmentation points have located based on the white color (background pixel) or the distance between two successive local maxima and one local minima with almost no foreground pixels. Figure 9 illustrates technique of extracting the words images based on vertical histogram.



Fig. 8. Extracting the lines images based on horizontal histogram

Likewise, extracting the characters images technique uses also the vertical histogram, which has calculated based on the distance between the top and bottom of foreground pixels for the word image after thinning. Extracting of the characters from the words is required to remove the punctuation marks (dots). The dots here consider a major obstacle to identify the correct segmentation points of the characters. After determining the segmentation points, the dots will recover. Figure 10 illustrates technique of extracting the word images based on vertical histogram before thinning and Figure 11 after thinning.



Fig. 9. Extracting the words images based on horizontal histogram before thinning



Fig. 10. Extracting the word images based on horizontal histogram after thinning

The following equations show how the histogram has calculated based on the total number of the foreground pixels.

 $pr(r_k) = \frac{n_k}{n} \tag{8}$

$$s_k = T(r_k) = \sum_{j=0}^k p_r(r_j) = \sum_{j=0}^k \frac{n_j}{n}$$
 (9)

where, K=0, 1,, L-1,

N is total number of pixels in the image, L is total number of possible grey levels in the image.

IV. EXPERIMENTAL RESULTS

As a result of all the previous steps, in addition to the final stage of verification processes, the first issue of ACDAR database is now available. ACDAR database contains 208 pages, 208 paragraphs, 2,969 lines, 32,890 words, and 158,872 characters, the database is divided into two sets one for training and the second for testing. The details of the first issue of ACDAR database after extraction and validation process is shown in Table III. Table IV summarizes a comparison between the results of some databases use the Arabic handwritten scripts. Diversified samples from the ACDAR database have published in ACDAR's website under this link http://www.acdar.org/DBsamples.php.

Figure 12 displays samples of handwritten paragraph that written by one person with its printed text, Figures 13 to 16 display samples of complete free handwritten paragraph. More samples for characters, words, and paragraphs see http://www.acdar.org/DBsamples.php.

Details / writers	Each writer	Training set by 51 writers	Testing set by 53 writers	Total (training and testing sets) 104 writers	Percentage from the original
Paragraphs	2	102	106	208	92.0%
Lines	Average 30	1,467	1,502	2,969	87.6%
Words	358	16,214	16,676	32,890	81.3%
Characters	1,916	78,584	80,288	158,872	73.4%

TABLE III. FINAL DATABASE AFTER EXTRACTING AND VALIDATING PROCESSES

TABLE IV. COMPARISON BETWEEN SOME OF HANDWRITTEN DATABASES USED ARABIC LETTERS

Database	Details		Year	Writers
Al-Isra [10]	 500 sentences 10,000 digits	37,000 words2,500 signatures	1999	500
AHDB [12]	 10,000 words for check p 	processing	2002	100
IFN/ENIT [8]	 26,459 Tunisian city nan 	nes	2002	411
Khedher and Abandah [21]	48 pages of text		2002	48
IFHCDB [22]	 52,380 characters 	 17,740 numerals 	2006	-
ADBase / MADBase [22]	 70,000 digits 		2007	700
Alamri et al. [24]	 11,375 words 46,800 digits strings 1,640 special symbols 	21,426 characters13,439 numerical	2008	328
On/Off LMCA [25]	 500 words 30,000 digits	 1,00,000 characters 	2008	55
Al Hamad et al. [2]	 20 Pages 500 words	40 paragraphs620 characters	2010	10
ACDAR – First issue	208 Paragraphs/Pages32,890 words	 2,969 Lines 158,872 characters 	2014	113

(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 3, 2015

handwritten	سونت جنوب نخليخ ويختطورعلمي مذارات يقعليم لعامي نو بلفت يضعاح ديد، نشريت وطعدجمسرة
printed	سوفت جنوب الخليج وبمنظور علمي وزارات التعليم العالي في المنطقة لضبط آلية التدريب وطرق جديدة

Fig. 11. Sample of part free handwritten paragraph

وتالى الله طعه با سم مديرة البراجي التعلميت في ما يتورمونى جنوب انخليع وبمنظور عامي غان شرك دوم حراساب الاتذا يتيت الذي أ بهيته مع وزارات التعليم في المنطقة سوى تتيتج لملات الرى الطولاب في العالم العربي ترجع اكبر لضط الدي الترمير وطرق جديره لعمل وربط مهاداته فى مجال تغذيات المعلمات دالاتصالات وربق عمليه تمناع عليها الاتفاعية بعا لا يتل ما يتكام قرص أكثر ارتباط بواسطت وربق عمليه تمناع عليها الاتفاعية بعا لا يتل ما عالمتي وربق، وض سعاد بالغمام مركب أسواج في العدير من الدل العربيت إلى عدن المبادر قرص أكثر ارتباط بواسطت مركب أسواج في العدير من الدل العربيت إلى عدن المبادرة من في المذام ما يتزومون المنظران المن عامل العربيت ولا عن ما تيت وربق، وض معاد با فعام مركب أسواج في العدير من الدل العربيت إلى عدن المبادرة من فن الذراح ما يتزومون المنظرين التي من شائلا أن تشرق تجرب التعليم وتعزز قدرة الملاب على التعلم بما المنظرين المي من شائلا أن تشرق تحرب المعاليم وتعزز قدرة الملاب على العلم بما المؤدمة أن معتني تردة العلم ما ليث أمرت عمل مرمك ومليخ المناجع السيره، راى ذلك أمن والما في المي من الما المين المي من ألما المرجع المعام على الما من مون مون والما مع المنا أن العالي مت لمون الما من وتعز الملاب عن المين من العز من من شائلا أن تشرق عدة المين من المين المي مرمك ومليخ الما جم السيوم المرة من والغاد في من المي من المي من المن الما مي وتعز في المناجع السيوم، راى ذلك أمن والما و من المي من المن الما من المي عيمن خليل في المين من من المي من المي من المين من من من المي الما من من خليل الما من المي من الما من المن المي عيمن أن من المي عن من من المي من المي خليل المي من المي من الم

Fig. 12. Sample (a) of complete free handwritten paragraph 1

· تشكل مبا در: " مدكا م بن بنعام " الن ا الملنز ما بجروم من عاطيًا ب أبلعل رلبلا ما بنمًا عل ا دراك ما يحدم من لاً ج النتیام التنابخ با المدارس ، سریدن قدمتد المبا در ۲۰ البُ تطبیع علیک بمنطعہ حالمی استفارا تا برجرین جا متر رمکا تقدیت مقبل متبتر امی ما نتین و مجتسبت ملبون رولار مل مدی السندات النمست بلقبلت لندريب ولفلابه مر المعكبت مل مؤرا ن تعتبت المعلدمان ملانعالان بد الدع الثقني اللازم فهم · متعقد الحبا در» دلغلا ي مطعكين ابغنا مجعف منابد فيها ن ما بجرم مدفت التعليب الخاجة سطر لي - با معار متخفظة . ونم نطيبة جد المبادرة بالنعامان بع ج بجدمة مع مسان م منعمان فاجم عا ينتاجو ب اجتبامات المبقعان الملب لكل بلد . من جهت خا ل عبد بلطيف محد مدبد كام ما يجدم من بالما بخلي الذ ما يجد محد أ فذن على فالغري مهية عكيت المدارس من المستقار مذاجا رسا تد تغنية بعليمان إحلب السلجية ، وعليه طعدت الشكة مبادرة " مشكراء من لنه موتم منا بيدكيبيدين من الحكومان مرتب مباور و محيوة مرتبعه موتم منا بيدكيبيدين من الحكومان مرتب مسبب ت الشائبة ميكرة مل الهجب الاستقلار المستراجل من لتفعنيات ميداج بشديب لن متدور مدرة لفلاب مل تهل ليلم حد لمعاكيت على نبطويد مركزاتهم معدر مدرم بعد با من مهل مع موصاحین من موجد مرز مرز مرز من منها الما دا سالب التدریت - دا شرار ای آن مبا درم " ستی و ن مشله" الن تشکل از جد معا مل المجاح الا ساسب مل بجدد سرمت ما الم مساعدة الاندا و - المد مسسان ر مشکن و مل استقار الم مساعدة الاندا و - المد مسسان ر مشکن و مل استقار جا مائم را الا منة الا عماد عل لا را ن م لتعشيا ن رائي مدمها · cireren a

Fig. 13. Sample (a) of complete free handwritten paragraph 2

ومالك الناطي باسم عورة البرامي التناليه في ما يكرومرن جنب المنابع و يستخدر علمي خان متحدل مرض إساس الانتائي الني أ برمت مع مزارت المنابع في المنظنة سون تتسع كمان المحال الطلاب في العالم العزي من المحر أكبر نصل ألي التترسب وطرق حديث لمحفل مديط وبالأتي في مجال تقتيدات المعلمات والاترت مراحا مراحا مع من علم ومنم حمارين عن لمالها والمحار فرص أكم الركباط بواسل مدين معركة أطاع عز العدير من الدل العرب إلى عزه المبادر من المحال التناب حدث علم والمحم العدير من الدل العرب إلى عزه المبادر من المناح والمناح والمحل المحلوم المراح الدل العرب إلى عزه المبادر من المناح والمناح والمحل المحلوم المح المن الدل العرب المحمد عزل عمرار بالمن معركة أطاع عز العدير من المنظق ومع حديث المعلم من المحال المحلوم المح من المعام ومن والمحل المحلوم الم ما حاليا أن تركيب تحريف المحل مركز عنه العلام مع العالم والمحف المحلوم المن و عن حلي المحروب المحلي مكمة من العال محل والمحرف المحلوم المراح و محقيق من المحدوم التي بالمعام محض وحل العالم والمحرف المحلوم المن و محقيق من المحدوم المحل معلم محض وحل المحل والمح مراحي ما و محقيات من المحد العلم مركز عنه العلام عالم والمح من و محقيات مع المحدوم المحل محل محل محل ومحل والمحل المحل المح و محقيات مع المحدوم المحل محل محل محل والمح والمحل المحل المح و محقيات مع المحدوم المحل محم محل المحل والمح والمح من مع و محقيات محد المحل من الماح مح محل المحل المحلي المحل المح ما المحد المح والمح من المحد المحي على أن المح محل محل المحلي المحل المحل المحلي المحل من المحلي المحل والم

Fig. 14. Sample (b) of complete free handwritten paragraph 1

دليلاً واضعاً مم ارداك ما يكرر ونم لذهب تو تعلم اكتفاع الدارس ، و حف توفر ا بدردة النم تضم معلياً ينهز حلي ا مشمارات برمعية على ومنعاً فقر توفر ا بدردة النم تضم معلياً ينهز حلي ا مشمارات برمعية على ومنعاً فقري تعلم عيميا ال ما لمتين و عنون مليون دولا مع مدى المناك الأسا المقبلة لتدريم العلام لمم ، وتوفر المداررة للطلاب والمعلين اريئاً مجوعة مس برمعيات عابكرو وفق المعادرة على عارات تعنية والعامين اليئاً مجوعة مس برمعيات عنابكرو وفق المعادرة على عارات تعنية وعار معاد من تعمير مدة المدرة با لعادن مع عبات على عاملة والمعلين اليئاً مجوعة مس برمعيات مدة المدرة با لعادن مع عبات على عادت وما حمل ومنفات ما معن ما مدة المدرة با لعادن مع عبات تكومت وعة مسان ومنفات ما محمد ما مدة المدرة با مقادن مع عبات المعاد والعالي المد معموض ما معت ما مدة المدرة عارات المعتمات المعلية النا ما من معان ما قد مع ما ما يكرد وفق حب والحلي ان عايكرد مع ما من عامت ما من عام مايكرد وفق حب والحليم ان مايكرد مع ما من عامت ما مدة عليه مدير من ما عارات والما من ا من مرابع وهوا معان وما معارات و تأسيد كبيرين مع ولكومات والما من ابن ومنا من ومارات من الما و تأسيد كبيرين مع ولكومات والما من ا من الما من الما من الما و ما مع مع ماراتم و أساله منه المام المتوافي معاراتم و أسال المام المام المكرد مع المام المن المام مع مع ما من عاراتم و أسال المام من المام من المار من العام و ما مع مرد عام الرام و أساليه المام من الما مع من منه المام مع مع مع مع منه المام و معاراتم و أساليه المام من المام المام مع مع منا من منه المارة منه مع معام المام المنه المام المام المام مع مع منه منه المام منه معاراتم و أساليه المام من المام معان منه العام مع مع مع منه منه المام المام و معاراتم و أساليه المام من المام منه المام المام مع مع منه منه المام المام و معاراتم و أساليه المام مع منه ما مع منه منه العام م مع منه منه منه المام منه المام منه المام مع منه منه منه منه المام المام منه المام منه و منه منه منه المام منه المام م ولما ما يكرد وم منه .	و تشكل مبادرة " شركاد ع المعلم " ان الملقها مايكرد موقة عالمياً ع ألمول
حَوْمُ المبادرة التي تَضْه معلياً ينها و المتهارات برومية مهدة ومنعاً تقريب دَمْل عَتْمَها ال ما لمنتي و هـ فن حليون دوفار على عدى السنوان الأسا المقبلة للترب المطلاب والمعني على معارات تقنية المعلومات والارتبالان وتوفير المدم المتقيق اللازم لمم . وتوفر المبادرة للطلاب والمعلية ارتبا ومجوعة مس برمعيات حمدة المبادرة بالمقادن ع عهات عكومية وعالمات ومنفات . ويتم تطبير معا يكرد وفق المبادرة للطلاب والمعلية ارتبا وعنفات . ويتم تطبير حمدة المبادرة بالمقادن ع عهات عكومية وعالمات ومنفات عامت مما حمدة المبادرة بالمقادن ع عهات عكومية وعالم سان ومنفات عامت ما عند عدم عام عايكرو وفق حبات المعلية لك بلد . مسرعيت قال عمالطي قد عدم عام عايكرو وفق حبو الخليج ان عايكرو حقت أخذ تقنية المعلومان عن العلمة تعكيف المدار سا مد استمار عزايا وحوائد تقنية المعلومان عن العلمة المعلي ، وعليه طورت المركمة عبادرة "سركادع المعلومان وتأسير كبيري عد الكومات والمؤسس المتيات ومناة عنه المتواهل ع المتنات و مراجع المديمان المناح المالات عامرة والمعلين على تضور معاراتم وأساله المناح الأسلة على المالا المعام وتأسير كبير عام الرواجة المارية المالية والمعان عالمان المعام والمعلين على منها المن المالية المعام المناح المن عن المعاد المعام والمعلين على تضور معاراتم وأساله المناح الأسلية على المالية المعار المعلين عان ما معاد المارة المالية المارة المناح المارية على المارة المارة والمعادية عالمان معاراتم وأسالية المناح عليه عالمان المارين عن الماليم المعولي ما ما معار من المارية المناح المناح المن عاملة المارة المارة المارة المعادين عام المارة من المناح المارية والمانة المارية عارائم وأسل المناح المارية والمالية المعولي عان ما من المن المارة وأسالية المارية والمانة المارية عارائم وأسل المناح المناح المارة المارية والمعان المارية والمارة المارة المارة المارية المارة المارة المعولي مانا مارد والمارة المارة المارية والمانة المارة عالمانة المارة والمارة والمارة والمارة المارة والمارة المارة والمارة والم	دليرا واضعا عيم ادراك مايكرو وفت للهميت معلم اكتفنا ع المدارس ، وسعف
دَمِن عَسَمًا ال ما لَنَيْ و عَن ما يون دولا عا مدى السواق الذي المقبلة لتدريم المطلاب والمعاين على مهارات تقنية المعلومات والانهالان و تونير الدكم المتقيم اللازم المم . و وَفر المباررة للطلاب والمعلية ارينا و موفق هسر برمعيان عاركرو وفق المتعادن ع وبات عكومية وعلى حسات ومنظات عامته ما هذه المبارة بالمعادن ع وبات عكومية وعلى حسات ومنظات عامته ما تيناب ع ا عيكرو وفق حب الملية المعلية الك بلا . مسرعيت قال عسالطية قد عدم عام عايكرو وفق حب الملية والمعلية المعار وفق من العالية معة وكدين الملات المعمينات المعلية الك بلا . مسرعيت قال عسالطية معة وكدين الملات وتعامل من المعار من الما ومنا معارك و توفق المعالية و تأسير كبيرين مسر المحكومات والمع حدات المعاركة عامتها المع و تأسير كبيرين مسر المكومات والمؤسسان المتناف مركزة عام ألهية الستمار و المعلين على تضور معاراتم و أسوا المناح الما عمارة المعار و المعلين عان تضور معاراتم و أسوا المناح الما عن الماره من الما مع المع الما ع المتناك و براعة المعاركة و المعار الما مع المارة المحلول تأني ما مه الرؤية الا ميات المناح الما عن الماره المارة المع من عام المارة و الما عن المن عاد الما المناح الما المن الما المار والمع من عام المارة و أسوا المارة و الما المناح الما المارة الموارة المعامين عام المارة م الما من المناح المارة المارة من المارة المارة و والمع من عام المارة و أسوا المناح الما من من الما مع من من الما معار و مع لم عارك من منا المارة معام المناح المناح عار المارة المارة المع من ما عارة من الما من المناح المن عالمان ما من المن من من المالية الم المع من عام المن المارة ما المناح المناح المن عام المناح المن من المارة المارة و دا مركن عا استمار الما ما من المانة عام المانة الما منها عالمان من الما من مراح المارة المارة المارة المارة المارة من المارة من المارة من المارة من المارة من من من المارة والمارة المارة المارة و المارة من المان من مرد من منه من منه من المارة المارة و من من من منه المارة من من من منه منه من من من منه من من منه المارة من من منه منه من من من منه من من منه من منه من من من منه منه	توفر المبادرة النم تضبه وعلما فينظور عالمي استثمارات بروجيت علمت وعنعا فقريص
لتدري المطلاب والمعلمين علم مهارات تقنيت المعلومات والاشلاف وتوفير المدم التقني اللازم المم . وتوفر المبادرة للطلاب والمعلمين ارنية وجودة مسر برمعيات عاركرو وفت المعادن علا وممات حيط الماتية وبا سعار وضفت . وبتم تمسير هذه المبادرة بالمعادن علا مهات عكومية وعال سمان ومنظات طاماته بما مدم عام عاركرو وفت حوبات عكومية وعال لمل . مسرعيت قال عسرالكي قد صبر عام عاركرو وفت حوب الحليج ان عاركرو معت أخذ تحال عارتهما معهة تمكين المدار سما مد استمار مزايا وها من تقريت عامتهما عن العلية المعلمين ، وعليه حوارت المركمة حبا درة "سركادع العلومان عن العلية العلي ، وعليه حودت المركمة حبا درة "سركادة المعلومان وتأمير كبيرن مد ولكومات والمؤسسة التركية عارز مرابا وها محملة العلومان وتأمير كبيرن مد ولكومات والمؤسسة المركة ما درة "سركادة العلومان وتأمير كبيرن مد ولكومات والمؤسسة مرابا ومرابا العام والمعلمين على تضور معاراتهم وأسل المناح الأسلم عرابا معارما محمول تأني من علما المراح العد عالي المناح المالي من المحمل المحمول تأني من علم الردية العراقيم والماتها و المواح على المارة دو المعارية عالمان الما عام المعام المناح الما علمان من المعام المحمول تأني من علم الردية الماتيم والمانة عاما والمان عارزة المارة دو المعارة عام المواحة المعام المناح المناح المواحة المعام الماتين المحمل المناح المان من الماتهم المات المناح المان عامرة المعام المعام المحمل الماتي المحمل المالية الماتي المات المات المارية المالية الماتهم المان المالي المان المان المان المان المان المناح المان معادرة المان المان المان المان المان المان المان المان المالية المان المان المان المان المان المحمل المان مع المان المالية المان	دَمِن عَسَمًا ال ما نُسَى و 2 ف مايون دولار على عدى السوان المنس المقبلة
التقني اللازم لمم . وتون المدادرة للطلاب والمعلين ارينة معجودة مسر برمبيات حايكرو-ونت المتعليد الخاصة جطع المكتب وبالسعار حفظت . ويتم تتهلير حدة المدادرة بالمعادف مع جهات عكومية وعو مست ومنطات طاعته مما تيناسب مع ا عيتاجات المصبحات المعلية لكل بلر . مسهمة عال مسالي قد عدم عام عايكرو-ونت حبوب الخليج ان عايكروسفت أخذ تحك عامقها معهد حكين الملارس عد استثمار حزايا وحوائد تقنية المعلومان معهد حكين الملارس عد استثمار حزايا وحوائد تقنية المعلومان في العلية المعكيت ، وعليه طورت المركمة حبا درة "مركادع المعلومان وتأسير كبيرين عد المكرمات والمؤسسات المتعانية حركرة عام أهية السمار وتأسير كبيرين عد المكرمات والمؤسسات المتعانية حركة عامة العام وتأسير كبيرين عد المكرمات والمؤسسات المعادة المعلومان والمعلين على تضوير حياراتم وأسايد المتواج التياج الألف معاد م ومحول تأنيا من علما الردية المعارة على دواعلة المناع المن عادة المحلين عالي من المار م المعاد المعارة م وأسايد المناح الالانيا المارة المارة المعاد عالي من علما المراح المعام المناح الماح على المارة المحاد المارة المعاد عالم المارة مع الماح الماح الماح على المارة المعام المارة المعاد عان المعام المارة وأساية المياد عالية المحاد المارة المعاد عالية من علمان الماح الماح الماح المارة عام المارة المعاد عالية عامية عاراتهم والماح عامة المارة المياح المارة المارة المارة المعاد عالية من عدم المارة معالية من عليا المارة المارة المارة المارة المارة المعاد عالية من عليا الردية الماح المارة عاد المارة المارة المارة المارة المارة المعاد عالية من عامة المارة عام المارة معاد المارة المارة المارة المارة المعاد عام المارة من عامة المارة عاد المارة عاد المارة من عادة المارة الما	لتدرب المضارب والمعمين علم ويارات تقنيت المعلومان والايضالان وتوفير اكدكم
ما يكرو وفق المتعليق الماجة ديلع المكتم وبال عار هفضة . و شم تطبير حدة المبادرة بالمعادف مع جهات عكومية وه فسب ومنفات عاجت مما تيناب مع ا ميتاجات المصبحات المصلية لكل بلا . مدهية عال عدال على قد صدر عام عاليكرو وفق ضو ب الحليج ان عايكرو مقت أخذ ترحك عائقها معهة تحكيف الملارس عد استمار حزايا وخوائم تتشت المعلومات مع العلية المعلية ، وعليه طورت المركمة حبا درة "مركاد في المعكم وتأسير كبيرين عد المكومات والمؤسب التي تعزز خارة المحراب عن لما وتأسير كبيرين عد المكومات والمؤسب التي تعزز خارة المحراب مثال والمعلي على تطوير عباراتهم وأسب التي تعزز خارة المحراب عن لما والمعلي على تطوير عباراتهم وأسب التيمات والم المعام المحول تأني من علم الرؤية الاستراتيم للاكرة والهاذي المادة المحول تاني من عد المحد الاستراتيم وأسلام الديا وعليه عارتهم وأسلام والم المعام ومحول تاني من عد الرؤية الاستراتيم المنها المعام والمعلي عارتهم المحد المحد على المراب على الم المن المعام وعلي على المعارة من المحد المحد المعام المنها المعام وعلي عاريم عليه الرؤية الاستراتيم والمعام الديا المحاد الم والمعان عار المعام الموات المحد المحد المحد على المات المعام المعام والمعام عار مع الما عام المات الموات المحد المات المعام المعام المعام المعام المعام المعام المعام المعام المعام وعلي مع المحد المحد المحد المحد المات المراب على المرادة الم المعام المحمول تاني من عالم المحد المحد المحد المحمد المنام المعام المحمد من عالم المعام المعام المحد	اكتقما اللازم لمهم . ودّونر المبادرة للطراب والمعلمين امضةً فجوعة فسر برفضات
هذة المبدرة بالمعادن مع فيهات عكومة وعد مست ومنفات عامت ما تيناب مع ا ميتاجات المصبحات المصلية لكل بلد . مسمهت قال عدالطيه قد صبر عام عاليكروسونة حب الخليج ان عايكردست أخذت ما عامتها فعهة تحكيف المدارس مد استمار هزايا وخوائد تقتنت المعلومان في العلية المعلية ، وعليه طورت المرتحة حبا درة "مركادي المعلومان وتأيير كبيرين مد المكومات والمؤسسات المشتافية وكرة عام أهية السمار المعلين على تضوير عباراتهم وأسابيه المتديمة والمشار المحالم المعام والمعلين على تضوير عباراتهم وأسابيه المتديمة والمارة المحارة "مركادي المعام والمعلين على تضوير عباراتهم وأسابية المتراجع المار المحام المحول تأنيا من علمه الرؤية الاستراتيم والمناع المراحة المحادة المحول تأنيا من علمه الرؤية الاستراتيم والمعام والمعان المراحة المحارة والمعان عار من عام المحارة من المحام المعام والمعام المراحة المعارة المراحة المحادة المحول عام المروحة مع المحام المعام والمات المحاد عام المحادة المحادة والمعان عار محمد المحدة المحد المحمد المحمد عام المحمد المحمد المحد المعام وعليه عن المحد المحدة المحد المحد على المراحة على المحرد المعام وعليه عاراتهم والماستراتيم والمن عادة المحمد المحد المحد المحمد المحمد المحمد المحمد المحمد المحمد المحمد المحمد المحد المحد المحد المحمد المحد المحمد المحد المحمد المحد المحد المحد المحمد المحد المحد المحد المحد المحد المحمد المحد محد المحد	مايكرد وفق التعلية الخاصة حط المكتب وبالسعار مفقفة . وتم تحص
تيناب مع ا ميتاجات المعتمات المعلية لكل بلا . مدجهة قال عد الطيف قد مدير عام مايكروسونة حبة مو الحديم ال عايكروسونة أخذ صحا عاتقها مهة قد كدين عام مايكروسونة حبة مو الحديم ال عايكروسونة أخذ صحا عاتقها مهة قد كدين المعلية ، وعليه حورت المركمة موادرة "مركادع الكلم" بدكم وتأيير كبيرين عد الكومات والمؤسسان المشاخة مركزة عام ألهية الله تشار المدوم على رامتنيات وبراهم وأسابيه المدرية . وأسار الماكن معادة ا محادي المعلية على تصور معاداتهم وأساب المدرية والهائم الماكر محلول تأني من على أحد عاداتهم والكنيم الدينة والهاؤة المحادة المعلم محلول تأني من على أحد المحد عام الدينة المعام محلول تأني من علم الردية الاستراتيم للاكنة والهاؤة المحروفة على المك معلول تأني من علم الردية المنت المكرة والهاؤة المعاد من المن المعاد الم م حلول تأني من عام المرادية المحد المحد على المؤمنة المحدة الفرادة دا سركان عا المرد من المعام الكان المحد عام المعام والمات المعام الم دا مركون عار المعام الماكرة المحد المعام المعام المنا المعام المعام المعام المعام المعام المعام المعام المعام المحد المحد المحد المحد المحد المحد المعام المحلين المعام الممام الممام المعام المعام المعام المعام	هذة المبادرة بالتعاون مع فيهات عكومية وعؤ سساق ومنفات فناجته بما
قد حدم عام عايكروسونة حبة ب الخليج ان عايكروسونة أخذ ترجل عائمة ما مهة حكيف الملاار مع حد استمار حزايا وحوائد تشنة المعلومات ع العلية للتحليف ، وعليه طورت المرمحة حبادرة "مركادي المتكم" بدتم وتابير كبيرين عد الكومات والمؤسسات للمتانية وكرة على أجهة الاستمار المتواجل ع المتنايات وبراهج الشريب التي تعزز قدرة المحراب على نهل العلم والمعلي على تطوير عباراتهم وأسايه المدس . وأشار الماكن مبادرة المحول تماني من جلب الرؤية الاسترتيم الدياح الأسلية عاليكروسوفته على الملاء وحلول تماني ما جلب الرؤية الاسترتيم الكراب علم المادة المحرار الع وعلول تماني من جلب الرؤية الاسترتيم الكران المار علم المادة المحول تماني ما جلب الرؤية الاسترتيم المكرة والهاد علم المراد المسلو والمعلية عالية من جلب الرؤية المحارية والكراد عام المادة المشار الن وحلول تماني ما جلب الرؤية المحارية والكرانة عام المنام الأ	يتناب ع ا عيّاجات المعمّعات المعلية لكل بله . مسجهت عال مسالل
مهمة حكيف المدارسوا مد استمار حزايا وموائد تتنبة المعلومات ع: العلية المعلية ، وعليه طورت المسركة حدادة "سركادي (ليكم" بدكم وتأيير كبيرين عد الكومات والمؤسسان المتنافية مركزة على أهمية الاستمار المعلومي على تطوير معاراته وأساليه المتدس . وأشار المحالب على مهل العام "شركا ديخ المقلم" التي تشكل أهد عوامل النعاع الأسلية ماكرة وقت على المعلى وحكولي تأنيا من عهد الرؤية الاسترتيمي للتركة والهادخة العسادة المعارد الم ومكولي تأنيا من عهد الرؤية الاسترتيمي المتكان الأسلية القارد الخسان والمعلومية على المترار علما تامم المكان علم المادة المعارد المسان وعلولي تأنيا من عهد الرؤية الاسترتيمي المتركة والهادخة العام العالم العالم والمعلومية عام عامة المحاتاتهم المكانة على المؤمون على المعلمة المعارد المعار المعار المعار المعالمة المعام الم	قد عدم عام مايكروسوفة حدة مما كليج ان مايكردسفت أخذت كا عاتقها
ع العلية (مقايية) وعليه طورت المسركة حبادرة "سركادي (لتظلم " بدتم وتأسير كبيرين عد وكترمات والمؤسسان (ممتافية وترزة عام ألم يترالاستقار المتواصل ع (متنشيات وبراح المتدرية التي تعزز قداة السلمات على نهل العام والمعلين على تطوير معاداتهم وأساليه المتدرية . وأشار المالي على نهل العام " شركاد ع (مقام" التي تشكل أهد عوامل النعاع الأسلية عاليكرد موقت على بلدى ومحول تأنيا من حلب الرؤية الاستراتيم والازفة الاصادرة الأطراد ال دا سركان عا استثمار حاقاتهم والكارت بالاعتماد عا الأمعاد والتقضات التي تو فرحا ما يكرد موفت .	مهت قيكين المدارسا صرا ستثمار حزايا وعوائد تقنيت المعلومان
وتأيير كبيرين صد وتكومات والمؤسسان المشتانية مركزة عا أهمية الاستقار المتواجل ع المتتنيات وبراحيم التدريب التي تعزز قدرة المطراب عى نهل العلم والمعلمين على تضور معاراتهم وأسايد التدريب . وأسار المان صادرة "شكاديع التقلم" التي شكل أهد عوامل المنعاح الأسلية عليكردسوفته على بلدى وصحول تأنيا من جلب الرؤية الاستراتيميو للتركة والهادفة الأصادرة الأسل دا مسركات ما استشار طاقاتهم ولكانت بالاعتماد على الأمعان والتقنيات التي توفيهم ما يكردسوفت .	عُ الملية (تعلية) وعليه مورت الشركة صادرة "شركادع العظم" بدئم
المتواجل ع المتقابات و مراجع المشرب التي تعزز خداة المطراب عن مهل العلم والمعلين على تضوير عادارتهم و أساليه المتدسيد . و أسار الكافن مبادر ة " شكاريخ المقلم" التي شكل أهد عوامل الذجاح الأسلية عليكروسوفت على بلدك الصحول تأنيا من جلب الرؤيت الاستراتيمي للتركة والهاذهق الأصادرة الأسل دا مسكول عاني الشمار طاقاتهم ولكانت بالماعتماد على الدوحات والمتقنيات التي تو فرحا ما يكروسوفت .	وتأسير كبيرين عد الكومات والمؤ مسات اكتقافية مركزة عا أهمية الاستقار
والمعلمية على تضوير عباراتهم وأساليه المترسي وأشار الاأن مبادرة " " سكادين المظلم" التي تشكل أهد عواهل الدنباح الأسلية للايكرسوفت على المدى الصلح لى تمانيا من عليه الرؤيت الاستراتيمي للاكمة والهاذه الأصادة الأخرادة المسملة والسيكاف مع استثمار طاقاتهم والكانت بالاعتماد على الدوهات والتقنيات التي توفرها ما يكروسوفت .	المتواجل ع التقنيات وبراجع الشرب التي تعزز عددة الصراب مى تهل العلم
" سَمَاد في النظام" التي تَشْكل أُهد عوامل المنجاح الأسلية لليكروسوفت على بلاك الصحول تأنيا من حلب الرؤيت الاستراتيمي للشركة والهاديف الأصادة الأخرادة المحسمان دا مُسْكِاف ما استثمار طاقاتهم المكانت بالماعماد على الدُّمعان والتقنيات التي توفرها ما يكروسوفت .	والمعلين على تضوير حياراتهم وأكاليه اكترب وأكرار الاأف صادرة
الصحولي تماني من جلب الرؤيت الماستراتيمي للشركة والهاذيف الأحسادة الأخرادة المسمل دا مستمانة مح استثمار طاقاتهم والكانت بالاعقاد على المد حوال والتقنيات التي توفرها ما يكروسوفتة .	" مركادي العام" التي تشكل أحد عوامل اكتباح الأسلية ما يكروسوفت على المدى
دا بشركات ع استثمار طاحًا تمم الكانت بالماعتماد ع الأمعان والتقنيان التي توفرها ما يكروسوفتة .	وصفيل تأيا ما عنه الرؤية الاستراشي للشركة والهادعة الاصادة الأفراد الوسان
دَو فرمها ما يكرو - وفت .	دا ب كان ما استمار طاقاتهم الكادف المعماد ما الأموان والتقنيان الم
	دَو فرمها ما يكرو وفت .

Fig. 15. Sample (b) of complete free handwritten paragraph 2

Figures 17 and 18 show samples of free handwritten lines extracted from the paragraphs.

خصه اكبر فضل الني النترمي وطرق حديده لصفل مبابط وسلافت ه مجالى
فرصة أكبر لضبط آلية التدريب وطرق جديد لصقل وربط مهاراته في مجال
والاتصالان متونير الدع الثقني اللانع للطم . و توخر الملادع للطلاب مسالمقن
والاتصالات وتوفير الدعم التقني اللازم لهم وتوفر المبادرة للطلاب والمعلمين

Fig. 16. Samples of ACDAR free handwritten lines

عدم ما محوود وفت جنوب الخليج ان ما محرود وه الخذ i lien o the server and a Les ye colail, of the تعال قيعتها إى مانتين بخمسون سلبوه و ولاماه سرم السوان وفوالا تقليه المعلومات في العملية (يعليه القعلمية) وعلية اجذب على عالقها مهمية تعكس المدارس مد الستقار مزايا وموائد تصب المستالي فحس المقلحة لتبييني الطلاب والمعكسر على تصارات تضيقه المعلونات والماضلات وتعازرالدهم شعول ووضوح أساس الاتفاقية الذي أبرمته مع وزارات التعليم في العنطقة سوف تشيح المنات الرؤمة الاسترائيسة للشرى والهادن فالمساعدة الأفراد والمؤسسات والشريحات على استعشار وتعاز قدرة الطلاب على التعلم نهدا يستثم من وحول المعتمرك العلى متسلحين بالمبادئ لايقل عنه مانتي وردشهت ونخا سعداد بأ نفاء مشك مها ما تم وأسامه المددس واست الا أن عدادة "متركاء العلم" الى تسك العدي مدجة أكبد لفنبط آلبة الندريب وفدن جديدة لصغل مدسط ومواعاة طروف عليم عغير معارفهم من خلالعاء دابتكار فرص اكثر ارتباط بواسطه

Fig. 17. Samples of ACDAR free handwritten lines written by different persons

In addition, the first issue of the ACDAR database contains different samples of Arabic handwritten words; Figure 19 shows samples of handwritten words with their printed text, Figure 20 shows samples of free handwritten extracted from the lines image.

A sample of the characters that extracted from the words image and wrote by one person is shown in Figure 21; Figure

Handwritten	Printed	Handwritten	Printed
الذعي	الذي	ومكال	وقال
ددعم	ودعم	ومراعاة	ومراعاة
iqu	أحد	التعليم	التعليم
لدول	الدول	21-1	امواج
بطلاب	الطلاب	Un Lul	أساس
-12-	مساعدة	وزارات	وزارات
ارتباط	ارتباط	ما بنکار	وابتكار
Land T	باسم	وبخر	وغمر

22 shows samples of free handwritten characters wrote by many writers.

Fig. 18. Samples of ACDAR free handwritten words



Fig. 19. Samples of ACDAR free handwritten characters written by different persons

1	2	3	4	5		9	1	9	10	11	12	13	34	13	16
۱	ι	U	4	*	•	2	2	*	-	2	r	1	1	U	a
17	- 18	29	30	25	22	23	24	25	28	27	28	29	30	81	12
~	w	up	ve	R	P	þ	ie	4	لم	ح	2	ء	+	0	6
33	ы	35	16	57	38	39	40	41	42	43	44	45	45	47	48
4	-	78	2	0	0	í	ś	5	J	٦	٢	٢	٢	\$,
49	50	51	52	53	54	55	56	\$7	58	38	60	61	62	-	
u	~	e	a	2	P	2	2	U	v	2	3	L	L		

Fig. 20. Sample of ACDAR character that wrote by one person

1	1	-	2	*	*	8	8	3	+		1.0	1	2	115	U-	-		04	64	-2	1	4	
2	1	4	8	t	*	2	0	0	*		•	-	J	9	2	5	5	4	1	3	1	4	
	-	6	6	٥	٩	*		9	9	es	0	i	3	v	¥	1	2	0	e	-		2	
2	3	P	0	2	1	5	5	U		-	up	40	*	-1	4	4	4	7	8	5	3	-	
14 6.5	-		2	2		2	U	\$	5	5	3	7	2	7	7	2	4	0	~	0	*	7	
44	5	3	5	311	200	327	2	4	24	L	14	24	244	246	S	5	24	24	210	251	511	111	
2.56 (Jac	2.50	216	200	2.10 [2-0]	18	140	24	14	k	N	2	5	R	3	2	H	++	14 "E	***	541	Le	2	
14	5	114	57	10	1	100	-	10	-	210 2	14	10	1.0	220	7		2	10	204	111	2.00	10	
and a state	3.00	340	340	14	14	14	340	200	5	10		18	**	1.11	10	3.56	310	3.98	511		10	10	
10	1.52		-12		4	43	4.8	3	4.5		4.0	8.11	+11	4.11		4.0				***	1.0	10	
	12	4.38	4.5	10	10	10	10	+20	10	4.0		14.1	10		-10	4.8	1.0		-	14	10	-	
***		-	0.0			4.57	- 12	+10	414	411	410	-	4.12	*77		-	-	33	1	31	11	11	
14	C bit	53	H	10	511	511	3.0	5.14	111	310	5.02	1.11	311	5.00	340	34	5.00	4	340	120	F11	3.00	
50	2.	5.51	512		. 511	310	3.0	3.0	511	5	5	540	C.	340	344	30	5-0	1 4	54	10	530	5-77	
Q	*	D 550	9	3	S	5.55	1.00	140	10	U 340	1	L nz	0	4	*	*	2	2	*	410	2	***	
610	1	50	411	+0	W	40	90	10	*	4	4	40	10	2	5	10	-	1 9	10		2	60 +11	
0	10	0	1	5	L	5	1	3	1	4	-	4	0	~	5	A	R	A	2		537		
1	3	>	K Pre	1	1	24	6	*	3	2	2	3	P	3	-	1	1	-	-	M	-	4	
-	-	P	7	14	1.0	4	6	5	*	*				-		2	0	0	á	í	2	t	
1	3	٢	-	\$	-	4	~	2	0	2	P	2	-	u	4	2	3	L	Ц	1	١	U	
C	-	-	2	2	2	Ф	د	+		3	v	5	-	~	of	-	e	P	Þ	4	4	4	
-10	t		S	6	0		2	2	2	_1	1	5	5	3	3	1	J	5	~	÷	-0	U	
41	0	-	~	Ð	5	9	-5	10	1	14	1	N	3	l	2	-	~	-	2	3	3	10	
1	2	j	3	55H	+11 CP4	10	114	*** 00	518 C/E	141	141	1	5	i	1	14	2	e	5	11		4	
9	10	10	. 13	+0	in a second	517	1	**	**	527 J	10	10	***	*0 9	13 C	+0	+3 0	2	*** 8-	14	9	*** •	
100	15	18 1	100	53	2.00 V	10	34	14	14	241	24	2	2	14	10	10	-	100	19	8-54	1.55	**	
9.07	10	500	340		0.41	101	100	10.0	10.4	7	854	107	19-8	19.8	10-10	1511	19-12	38.00	10.14	10.15	8-4 6	16-17	
56.50	2.18	16-20	80	19.22	-	10.54	-	8-3	100	-	19.29	5.0	50	8-12	-	10.04			8.0		200	340	
50	0	10-40	104	100	24	100	19-42	040	10.00	0	19-12	10.11	19.54	10.55	0.9	1057	19.9	1 10 M	10-00	8	1142		

Fig. 21. Samples of ACDAR free handwritten characters written by different persons

V. ACDAR – A NEW CENTER

As any other center in this area, ACDAR [26] contains a set of internal sections describing the main objectives of the center and its contents, these sections cover many activities in the area of document analysis and recognition such as teach courses, research, publications, resources, people, contact details. While analysis of documents and handwriting recognition continues to be our primary interest, we propose research and software development projects involving diverse digital document types.

ACDAR is dedicated to re-build and re-structure a reliable and standard a benchmark database and set of integrated tools for handwritten Arabic scripts within, it is newly established a website (http://www.acdar.org). The website includes details about the center, and a sample of the first issue of the benchmark database. In the conceptual framework of ACDAR, many functions would give ACDAR its identity, mission, and direction. These centered on the benchmark database, research, training, and collaboration with the community. More details about ACDAR center see Al Hamad *et al* [26].

VI. CONCLUSION

The paper presents new techniques for extracting the first issue of a new benchmark database, it has written by 113 distinct writers with different ages, cultures, and genders. Two paragraphs cover all shapes of Arabic characters have scanned with different resolution; the final database contains 208 pages, 208 paragraphs, 2,969 lines, 32,890 words, and 158,872 characters. Half of the database assigns as training set; another part assigns as testing set. For extracting and validating the proposed database, the research has developed and tested a set of new techniques. An example of these techniques are pre-processing of the images such as thresholding, filtering, local minima and maxima of vertical and horizontal histogram for the segmentation, in addition, developing skew detection / correction technique, etc. The techniques have examined and tested through several experiments in order to use them later for creating a comprehensive database that we seek to cover all Arab countries. The paper also displays a comprehensive details of forming a new center for analysis and recognition Arabic handwritten scripts, the center calls "ACDAR". Functions and activities of the center have identified and explained in detail.

REFERENCES

- [1] Plamondon, R., S.N. Srihari, "On-line and Off-line Handwriting Recognition: A Comprehensive Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, 2000, pp. 63–84.
- [2] Al Hamad, H.A., R. Abu Zitar, "Development of an Efficient Neuralbased Segmentation Technique for Arabic Handwriting Recognition," Pattern Recognition, vol. 43(8), 2010, pp. 2773-2798.
- [3] Blumenstein M., "Intelligent Techniques for Handwriting Recognition. School of Information Technology," PhD Dissertation, Griffith University-Gold Coast Campus, Australia, 2000.
- [4] Lorigo, L., V. Govindaraju, "Off-line Arabic Handwriting Recognition: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28(5), 2006, pp. 712–724.
- [5] Hamid, A., R. Haraty, "A Neuro-Heuristic Approach for Segmenting Handwritten Arabic Text," ACS/IEEE International Conference on Computer Systems and Applications, AICCSA, vol.1, 2001, pp. 1–10.
- [6] Hull, J., "Database for Handwritten Text Recognition Research, Center of Excellence for Document Analysis and Recognition (CEDAR)," Department of Computer Science, State University of New York at Buffalo, Buffalo, New York, http://www.cedar.buffalo.edu/ Databases/CDROM1, 1993.
- [7] Kharma, N., M. Ahmed, R. Ward, "A New Comprehensive Database of Hand-written Arabic Words," Numbers and Signatures used for OCR Testing. IEEE Canadian Conference on Electrical and Computer Engineering, 1999, pp. 766-768.
- [8] Pechwitz, M. et al., "IFN/ENIT Database of Handwritten Arabic Words, Proc. of CIFED," 2002, pp. 129-136.
- [9] Nouh, A., A. Sultan, R. Tolba, "An Approach for Arabic Characters Recognition," J. Eng. Sci, vol. 6, 1980, pp. 185–191.
- [10] N. Kharma, M. Ahmed, R. Ward, "A new comprehensive database of handwritten Arabic words," numbers, and signatures used for OCR testing, Canadian Conference on Electrical and Computer Engineering, 1999, pp. 766–768.
- [11] Pechwitz Mario, et al, "IFN/ENIT Database of Handwritten Arabic Words," Institute for Communications Technology (IFN), Technical University Braunschweig, Germany, Ecole Nationale d'Ingénieur de Tunis (ENIT), BP 37 le Belvédère 1002, Tunis. IFN/ENIT, 2002, http://www.ifnenit.com/.
- [12] Alma'adeed, S., D. Elliman, C. A. Higgins, "A Database for Arabic

handwritten Text Recognition Research," Eighth International Workshop on Frontiers in Handwriting Recognition, 2002, pp. 485-489.

- [13] AI-Ohali, Y. M. Cheriet, and C. Suen, "Databases for Recognition of Handwritten Arabic Cheques," Pattern Recognition, vol. 36, 2003, pp. 111-121.
- [14] Haikal El Abed, et al, "Online Arabic Handwriting Recognition Competition," 10th International Conference on Document Analysis and Recognition ICDAR, 2009, DOI 10.1109/ICDAR.2009.284.
- [15] Fan, X., B. Verma, "Segmentation vs. Non-Segmentation Based Neural Techniques for Cursive Word Recognition," An Experimental Analysis International Journal of Computational Intelligence and Applications, vol. 2(4), 2002, pp. 377–384.
- [16] Al Hamad, H.A., "Over-segmentation of handwriting Arabic scripts using an efficient heuristic technique," IEEE International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), 2012, pp.180-185.
- [17] Al Hamad, H.A., "Neural-Based Segmentation Technique for Arabic Handwriting Scripts," WSCG 2013, 21st International Conference on Computer Graphics, Visualization and Computer Vision, indexed by Thomson Reuters/ISI-WoS, Czech, June, 2013.
- [18] Al Hamad, H.A., "Use an Efficient Neural Network to Improve the Arabic Handwriting Recognition," IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2013.
- [19] Nixon M. S. and Alberto S. Aguado, "Feature Extraction and Image Processing. Academic Press," 2008, pp. 88.
- [20] Richard O. Duda and Peter E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," ACM, vol. 15(1), 1972, pp. 11-15, doi:10.1145/36:1237.361242.
- [21] Khedher, M., Abandah, G., "Arabic character recognition using approximate stroke sequence," Arabic Language Resources and Evaluation - Status and Prospects Workshop, 3rd International Conference on Language Resources and Evaluation (LREC'02), 2002.
- [22] Mozaffari S., Faez k., Faradji F. Ziaratban M, Golzan S. M., "A comprehensive isolated Farsi/Arabic character database for handwritten OCR research," In Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR), 2006, pp. 385–389.
- [23] El-Sherif E., Abdelazeem S., "A two-stage system for Arabic handwritten digit recognition tested on a new large database," In Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition (AIPR'07), 2007, 237–242.
- [24] Alamri H., He C. L., Sue C. Y., "A new approach for segmentation and recognition of Arabic handwritten touching numeral pairs," Proceedings of the International Conference Computer Analysis of Images and Patterns (CAIP). Lecture Notes in Computer Science, vol. 5702, Springer, 2009, pp. 165–172.
- [25] Kherallah Monji , Elbaati A., El Abed H., Alimi A. M., "The on/off (LMCA) dual Arabic handwriting database," 11th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2008.
- [26] Al Hamad, H.A., Hamdi-Cherif A., "The Arabic Center for Document Analysis and Recognition (ACDAR) - Structure and Perspective," European Conference of COMPUTER SCIENCE (ECCS '12), 2012.