OCR-Based Electronic Documentation Management System

Khalaf S. Alkhalaf, Abdulelah I. Almishal, Anas O. Almahmoud, and Majed S. Alotaibi

***Abstract*—Optical character recognition (OCR) is one of the latest technologies adopted in a lot of areas such as management, business, criminal and social networks. It consists of recognizing image-based characters and transforming them to real digital character that can be editing, written and displayed. In this paper we will demonstrate our experience on utilizing OCR technology to recognize some key information in selected management documents in Arabic language. In addition, this paper will discover the literature about OCR, address some challenges and share some important lessons with suggested research ideas that can be conducted in the feature.**

***Index Terms*—Ocr, text-recognition, business automation, Arabic text-recognistion.**

I. Introduction

Optical character recognition (OCR) and text recognition applications are used commonly in today business as well as in research. The real value is the effort and time that can be reduced by utilizing this type of application. For example, to re-write a paper-based document, it will take about 5 minutes. However, it will consume about couple of milliseconds to do the same task with up to 95% accuracy rate. There are a lot of applications, coding libraries and commentarial software for OCR in international languages such as English, Chinese and other popular languages. But for Arabic language, there is lake of application in this area. Furthermore, unlike to other languages, Arabic characters system is complex in terms of different situations of the single character according the position in the word. In this paper we will demonstrate our work for project called "Electronic documentation Management System (EDMS). This research project including a prototype of utilizing OCR technology to identify some key information from the management documentation. This project has been developed as web application using asp.net technology. After processing some selected documents the final results was outstanding, it can be clearly seen that the accuracy exceeding 95% and the project works as it should be.

Manuscript received October 10, 2014; revised December 16, 2014.

Khalaf S. Alkhalaf and Abdulelah I. Almishal are with the Computer Science Department, College of Computer Sciences and Information, Qassim University, Qassim Saudi Arabia (e-mail: kalkhalaf@kacst.edu.sa, aalmishal@kacst.edu.sa).

Anas O. Almahmoud is with George Mason University, United States.

Majed S. Alotaibi is with King Abdulaziz City, on leave from University of Jordan, Saudi Arabia.

II. What Is OCR?

First of all we should spend a couple of paragraphs giving background information about OCR.

OCR is the stand of optical character recognition which is field of computer science that recognizing image-based text from photos and transforms it to real digital character. OCR works like human ability in the brain to recognize the letters, numbers and symbols. OCR can read both handwritten and printed text. The performance of OCR is directly related to quality of input documents and pictures [1].

The first occurrence of OCR technology was in 1929 by Gustav Tauschek as a patent in Germany. Followed by Handel who obtained a US patent as well.

The first commercial OCR product was introduced by Kurzweil Computer Products in 1978 and the first costumer was LexisNexis [2].

Then, the OCR became a leading technology in the software market with high sales and profits.

III. Related Work

Before showing our work, we have to explore some different experiences regarding the OCR in general and especially for Arabic language. The following discussion is showing OCR technology in the literature and what exactly they reach.

In reference [3] the author discusses and reviews the research papers in Arabic OCR especially in hardware and compare between only software method and hardware, and they said hardware is faster than software but the hardware method are untapped in many ways especially for Arabic OCR.

In reference [4] the authors present a new technique to recognize a character in document have non-homogeneous text lines, and the test it on 15 maps, and they compare the result with another product name ABBY, the new technique shows improvement more than 30% comparing with ABBY.

In reference [5] the author presents a new system to recognize the Urdu alphabet, and he succeeds and makes a new system that can recognize the Urdu with 98.3% percentage when he tests it on some documents with good quality and some with poor quality.

In reference [6] the authors present a new OCR system for Arabic based on segmentation technique and the result shows us the recognition accuracy 90%, and they present a new word algorithm segmentation that separate horizontally Arabic words.

In reference [7] new approach for Arabic OCR present for different size text and different lines, and this approach was

achieves 96% recognition accuracy.

In reference [8] the authors present a new method for Arabic and Farsi fonts recognition based on scale invariant feature transform, and the present system does not need to noise remove or any pre-process except on the low quality image, the result after testing 1400 text images was impressing it gives nearly 100% recognize.

In reference [9] they design and implement a new system that recognize Arabic character without pre-segmentation based on based on describing symbols in terms of shape primitives and using mathematical morphology operations, the result shows us 99.4% for noise-free text and 73% for scanned text.

In reference [10] the authors bring the old documents with noise and they study for try to enhance it. The exploratory study concluded that the best way to enhance the document is from changing the scanned brightness.

IV. OCR Document Management System

A. Problem and Purposed Solution

In our organization, there a lot of paper-based documents for different use and purpose. The problem arises when we want to retrieve some important information from thousands of documents, it will take a lot of time and effort.

To tackle this problem, we build an application to read the most important information from common management documents based on the nature of that information and the position of the information in the document. For example in Fig. 1 illustrate the nature of one of most used document which is related to business trip decision. It can be clearly seen that is the position of the information was predefined.

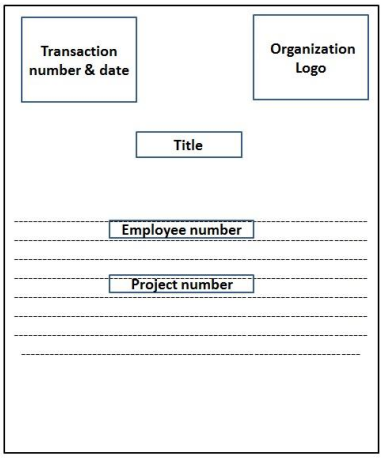

Fig. 1. Document structure.

B. Solution Structure

The software contains of simple web application that allows the department secretary to upload a PDF scanned document. Then, based on OCR engine, the software analyzes the document based on positions. After recognizing desired information, it will be shown as editable field to allow the user to correct some information that might be not well recognized. Then, the software saves the final information into Database for feature used. The Fig. 2 shows the general structure of whole solution.

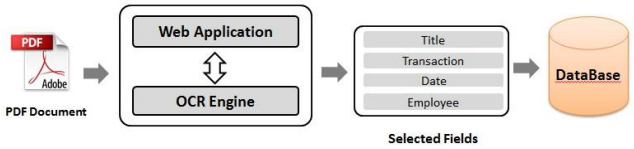

Fig. 2. Solution structure.

C. Main Features

Having like this kind of software, provide us some key advantages such as:

1) Saving a lot of management resources in terms of time and effort for the organization by reducing information retrieval process for some document and avoid losing some critical papers.
2) It is very beneficial for those organizations that still relay on paper work rather than automated management systems.
3) For those organizations want to transform to paperless environment, it is very powerful tools to use it for data migration process.
4) For Arabic language, we are still in the bingeing level. There is a lake of support and lake standards for this kind of software.

D. Main Challenges

Any kind of solutions faces some challenges and limitation and our solution is not an exception. We can summarize the challenges as following:

1) Arabic character nature

Unlike other languages, Arabic character is different because it has different situation according to its position in the word either first, middle, or at the end of the word. To clarify this issue Fig. 3 illustrates the different 4 situations of character (م) according to its position in the word.

This is considered as the main challenge of adopting OCR in Arabic language.

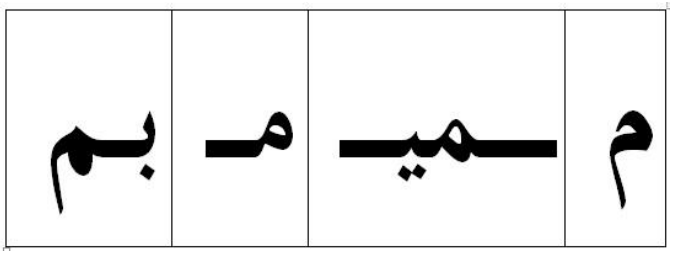

Fig. 3. Different situation of character.

2) Different fonts

Like other languages, the Arabic language has a huge number of different fonts. This issue will make the recognition process easier. However, this challenge can be addressed by unifying the font within the domain of use.

3) Image resolution

As we use different scanners from different users, it is difficult to use it as the same document in terms of quality of image. Some documents are in good condition, while other documents are in trouble situation. Dealing with these different types of qualities might be difficult process for both

user and the software itself. This will cause a verity of accuracy rates between same types of documents. The figure no.4 is showing the sample of scanned document.

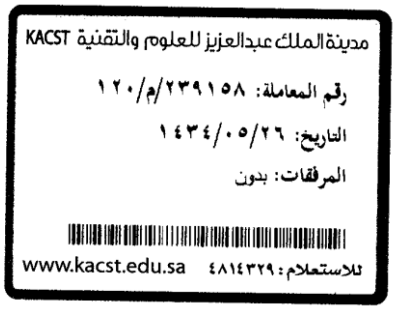

Fig. 4. Different situation of character.

4) Lake of experience

As stated before, for Arabic language we are still in the beginning level regarding this technology. However, some companies for example [11] try to take the lead and start developing outstanding software for Arabic text recognition. But compared with English language, we still looking forward for more products and researches that can be better for languages and our culture as well.

V. The Application

Our project is just a prototype, it not mature yet. Unlike commercial software, it is research-based software to enhance improvements in management process. For this reason, we apply the experiment for only one type of documents called "business trip document" to see the results and make some improvements before going to broadcast our experience.

The application is consists of the following parts:

1) The web-based application: This application contain 2 different user interface, first is for the department secritary to scan, upload and process the desired documents then save it to the database. Second is for the department manager to retrive all desired document with key words recognized by the OCR.
2) OCR engine: This is the core part of the application. It has been developed incorporation with Skhr Company [11] and utilized in the application as back-end service. The main advantage of this engine is the support of different fronts and the self –learning feature which allow the application to increase the efficiency as long as its work.
3) Database and reporting: Like any type of applications, our application stores the key information into the database in order to make the retrieval of the information and reporting easy. For example, if we want to know all trip decisions for specific employee, we can query as one click report.

VI. Results

After spending about 5 months for plan, design and implement the project, we got incredible results.

However, we still need more but we reach a reasonable level in terms of quality, efficiency and total cost. In our experience, we tried to scan about 4 documents that are similar to the following document:

Fig. 5. Sample of business trip document.

Then, we got the following results: Table I describing the document no. 1 and showing the differences between what is written in document and what is recognized from the application with corresponding accuracy rate.

TABLE I: Details for Document No. 1

Information	Original paper	After recognition	Accuracy rate
Transaction number	120/م/226140	120/م/226140	100%
Transaction date	28-7-1433	28-7-1433	100%
Employee number	4280543	4280543	100%
Project name	تحكم 2	تحكم 2	100%
Overall Accuracy	**100%**		

TABLE II: Details for Document No. 2

Information	Original paper	After recognition	Accuracy rate
Transaction number	120/م/228082	120/م/228082	100%
Transaction date	6-9-1433	6-9-1432	92%
Employee number	1428042	1428042	100%
Project name	امتياز	امتياز	100%
Overall Accuracy	**98%**		

TABLE III: Details for Document No. 3

Information	Original paper	After recognition	Accuracy rate
Transaction number	120/م/230694	120/م/230694	100%
Transaction date	10-11-1433	1-11-1433	92%
Employee number	4280945	4380945	92%
Project name	أفق	أفق	100%
Overall Accuracy	**96%**		

Sequentially, the Table II shows the results for scanning document no. 2 which also related to business trip decision document.

Table III and Table IV show the rest of the uploaded documents.

TABLE IV: DETAILS FOR DOCUMENT NO. 4

Information	Original paper	After recognition	Accuracy rate
Transaction number	120/م/233089	120/م/233089	100%
Transaction date	12-01-1434	0-1433	77%
Employee number	4310539	4310539	100%
Project name	كفاءة	كفاءة	100%
Overall Accuracy	**94.25%**		

To sum up, the following table will group all numbers and summarizes the overall rate.

TABLE V: THE SUMMARY

Document number	Accuracy rate
Document no.1	100%
Document no.2	98%
Document no.3	96%
Document no.4	94.25%
Overall Accuracy	**97%**

VII. LESSONS LEARNED

From our experience, we can share some important lessons that can be added to the OCR research area in general and for Arabic language especially:

1) Keep your software updated with new fonts and symbols or acronyms'.
2) Start from the latest technology without losing time to start from the beginning. For example, in our project we tried to develop our own engine but we faced a lot of difficulty to build the OCR begin from scratch. Finally, we reuse an existing powerful OCR engine and integrate our application with it.
3) It's important to organize your files to OCR engine rather than make it confused. For instance, save the time and mention the specific part of page to process rather than looking for all parts of the page.
4) As stated before, the main advantage of our OCR engine is the ability to learn while working. For example, if you are working in same work space containing a lot of documents, then you can let the OCR to tech itself by adding new characters.
5) According to the resolution of images, sometimes it's difficult to recognize the characters without errors.

VIII. CONCLUSION

From the above results part, it can be clearly seen that our proposed solution works properly with accepted Accuracy rate. We know that the numbers of tries are limited but as this is a prototype not a real implementation project, no need to spend more resource until we get a full support from our research center.

IX. FEATURE WORK

We are looking forward to invest more resources in this kind of software as it is beneficial in multiple sectors. This project needs a lot of improvement from different sides. Firstly, we need to increase the accuracy rate by supporting more fonts and include a lot if Arabic character situations. Secondly, we need to make a community of developers sharing the knowledge and Experience to increase awareness and utilize this kind of application in a lot of areas of applications. Finally, from scientific point of view, we need to innovate some new algorithms and techniques that ca be added to the field. Furthermore, to know the current practice in Arabic countries regarding this type of applications, we need to conduct empirical studies to know what is going on and analyze the needs of this kind of applications.

ACKNOWLEDGMENT

We would like to thank the national center of electronics and phonetics in King Abdulaziz City for technologies and sciences for direct support and allocating resources to complete this project.

REFERENCES

[1] R. Mithe, S. Indalka, and N. Divekar, "Optical character recognition," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 2, issue 1, March 2013

[2] P. K. Charles, V. Harish, and M. Swathi, "A review on the various techniques used for optical character recognition," *International Journal of Engineering Research and Applications*, vol. 2, issue 1, Jan.-Feb. 2012, pp. 659-662

[3] A. Beg, F. Ahmed, and P. Campbell, "Hybrid OCR techniques for cursive script languages – A review and applications," in *Proc. the Second International Conference on Computational Intelligence, Communication Systems and Networks*, pp. 101-105, 2010.

[4] Y. Y. Chiang and C. A. Knoblock, "Recognition of multi-oriented, multi-sized, and curved text," in *Proc. the International Conference on Document Analysis and Recognition*, pp. 1399- 1403, 2011.

[5] U. Pal and A. Sarkar, "Recognition of printed urdu script," presented at the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), 2003.

[6] A. Cheung, M. Bennamoun, and N.W. Bergmann, "An Arabic optical character recognition system using recognition-based segmentation," *Pattern Recognition*, pp. 215-233, 2001.

[7] A. Mesleh, A. Sharadqh, J. Al-Azzeh, M. Abu-Zaher, N. Al-Zabin, T. Jaber, A. Odeh, and Myssa'a Hasn, "An optical character recognition," *Contemporary Engineering Science*s, vol. 5, no. 11, pp. 521-529, 2012.

[8] M. Zahedi and S. Eslami, "Farsi/Arabic optical font recognition using SIFT features," *Procedia Computer Science*, pp. 1055-1059, 2001.

[9] B. Al-Badr and R. M. Haralick, "Segmentation-free word recognition with application to Arabic," in *Proc. the Third International Conference on Document Analysis and Recognition*, vol. 1, IEEE Comput. Soc. Press., Los Alamitos, CA, USA, pp. 355-359, 1995

[10] P. Herceg, B. Huyck, C. Johnson, L. V. Guilder, and A. Kundu, "Optimizing OCR accuracy for bi-tonal, noisy scans of degraded Arabic documents," *SPIE Proceedings*, vol. 5817, pp. 179-187, 2005.

[11] Sakhr. [Online]. Available: http://www.sakhr.com/index.php/en/solutions/ocr

Khalaf S. Alkhalaf graduated from Computer Science Department, College of Computer Sciences and Information, Qassim University, Qassim Saudi Arabia in 2009. He is currently studying master degree from the same department. The expected date to finalize the master degree is Jan. 2015.

He has total 5 years of experience in both public and private sector. This experience in software development for different platforms such as C++, C#, Java, enterprise

systems, oracle databases, Microsoft platforms and embedded development using C++. He also has an experience and training in software engineering and project management. Currently, he is working as a software engineering in King Abdulaziz City for sciences and technology. His job role includes development within different technologies to satisfy the need for software development in large research projects as well as managing small projects. He has attended different international conferences. He has two publications. His area of interests is OCR, data mining, big data, and software engineering.

Abdulelah Al-Mishal graduated from Information Systems Department, College of Computer Sciences and Information, King Saud University, Riyadh Saudi Arabia on 2009. He is currently studying master degree from the same department. The expected date to finalize the master degree is 2015.

He has total 5 years of experience in both public and private sector. This experience in software development for different platforms, such as Java, enterprise systems, oracle databases, Microsoft platforms and embedded development using C++. He also has an experience and training in software engineering and project management. Currently, he is working as a researcher in King Abdulaziz City for Sciences and Technology. His job role includes development within different technologies to satisfy the need for software development in large research projects as well as managing small projects for IT. He has attended different international conferences. He has 3 publications. His area of interests is enterprise applications, data mining, big data, software engineering, knowledgebase systems.

Anas Almahmoud received his undergraduate degree in information systems from King Saud University, Riyadh, Saudi Arabia. He has 4 years' work experience as a scientific researcher in King Abdulaziz City for Science and Technology. Now he is pursuing his master degree in applied technology specialized in cyber security at George Mason University, United States.

Majed Alotaibi was graduated from University of Jordan with bachelor degree in computer science in 2006. He has total 8 years of experience in both public and private sector and the experience in software development for different platforms, such as Java, C++, Oracle database, WPF, C#, Silverlight and asp.net. Currently, he is working as a software developer at King Abdulaziz City for science and technology. His job role includes software development, software testing and software engineering. His areas of interests are data analysis, software testing, database and mobile apps development.