P. Roochnik

ITISALAT OCR DISCUSSION

Announcing the Internet forum on Arabic computing, called: ITISALAT. ITISALAT, which made the pages of *Al-Hayat* (the world's most widely circulated Arabic language news daily) on 2 May 1995, promotes contact and stimulates the exchange of information in the field of Arabic computing. ITISALAT subscribers discuss the following topics and others:

Arabic optical character recognition (Arabic OCR);
Arabic computational linguistics;
Arabic machine translation;
Arabic computer-based training/education;
Arabic software development;
Computing developments in the Arab World;
Arabic corpora compilation;
Arabic data storage & retrieval;
Arabic hypertext;
Arabic code standardization;
Arabic email.

The name ITISALAT is actually a play on words. On the one hand it derives from the Arabic word "ittisaalaat" which means "connections" or "communications". But it also forms the acronym for *IT IS Arabic Language And Technology*. More than half the 225+ ITISALAT subscribers live in North America. The rest are spread throughout Europe and the Middle East, including Egypt, Iran, Israel, Kuwait, Morocco, Saudi Arabia, Tunisia, and the UAE. And although they hail from widely diverging professions, such as computer engineering, Arabic linguistics, and library science, they do share a common objective: to further the development of Arabic computing. ITISALAT began operating in May 1993.

The same wave of computer network communication that has captured the American imagination has begun to sweep over the Middle East. More and more Arab countries have connected to the Internet in recent years, and it won't be long before Arabic email becomes the medium of choice for communication with our friends and colleagues from Morocco in the West all the way to Iraq in the East. We hope that ITISALAT can play a small part in the Arabization of the Internet. Time will tell. ITISALAT frequently operates in cooperation with The Association for Arabic Computing in North America, ICEMCO, and The International Association for Arabic Computing. It has no official links with these or other organizations, but we all complement each other, and share a similar goal: to further the development of Arabic computing.

Arabic optical character recognition constitutes one of the most important topics of discussion on ITISALAT. On the surface, the discussion has concentrated on comparing and contrasting the various Arabic OCR software available on the market, along the dimensions of price, platform, ease of use, and quality of results. I hope that the readers of Manuscripta Orientalia will find the summary of ITISALAT discussion interesting*.

Even though I may seem an expert on Arabic OCR, I must stress I'm just an end-user with experience with two OCR-programs. I've reached acceptable results by using Al-Qari, but still sometimes (like yesterday night) I curse the product. I haven't published on the subject yet, but I'm now in the stage of preparing a paper about this subject for ICEMCO. This is also going to be a practical report on working with Al-Qari (and some remarks about the other program Iqra'). I will present some results during the presentation of the paper...

...I think I may say I am an experienced user of Al-Qari' al-Ali (Automatic Reader) of al-Alamiya (Sakhr), and also of Iqra' 4. Al-Qari is the newest and best performing of the two, so since about 6 months I stopped using Iqra'. I realized that in my enthusiasm when I reported on 99% success rate in Al Qari al-Ali , I maybe forgot to mention the spell-checking as being part of the process. So to repeat: I copy a page, remove all 'noise', scan it to make a -pcx file. Then I have Al-Qari recognize the pcx-file. This is done at an average speed of 80—90 characters per second. When the text-file is on the screen, I run the spell check utility, that has a relation with the picture-file. When the spell

* To subscribe, send the command (A) to address (B): (A) — subscribe ITISALAT your-1st-name your-last-name (B) — listserv@listserv.georgetown.edu. For more information, contact: Dr. Paul Roochnik, Moderator, ITISALAT email: roochnik@ios.com. Paul Roochnik.

checker runs into an unknown Arabic word, this word is both highlighted in the text-file (right half of the screen) and in the (enlarged) picture of the original (left side of the screen). In this way it is very easy to correct the misrecognized word. This may take some time though. I think an average of 8 pages an hour can be done in this way. But this work can be done by a young/cheap assistant, so that we researchers can lean back and think, or surf on the WWW. But if a misrecognized word is still an existing word, spell checker just proceeds assuming nothing went wrong. So 100% is not possible. I checked Abdel-Hadi's Home Page (for the OCR-report). Joseph Bell who reported on Al-Qari earlier this year is on this list too. I think he was less positive because the type of texts he is dealing with is completely different: old printing with all the disconnected characters etc. But Joseph also once wrote on this list that Al-Qari's capacity for storing ligatures is limited. I must say I did not run into this problem. First of all there is a number of pre-defined ligatures (all the usual ones, some 30 or 40 of them), but there is also the possibility of user-defined ligatures. I did not count them, but I think my most trained typeface contains at least another 50 or 60 ligatures (sometimes just combinations of 2 characters that are not ligatures at all, but the program always picks them together).

I admit this 99% (which is just an estimation) can only be achieved under most favorable conditions, i. e.: very good paper quality, very clear print, after copying the original and removing all noise (headers, footers, illustrations etc.) and learning the type-face for several hours (especially when the typeface contains a considerable number of ligatures). The number I mentioned was realized on the Kuwaiti magazine 'Al-Arabi' which reaches us through the Kuwaiti embassy.

Jan Hoogland

Now a brief comment on the Arabic OCR programs: I have followed the reports on some of the OCR packages available for Arabic and posted on this list or some other similar lists but never felt the need to make any comments because I did not have any access to such programs for the simple reason that I could afford none of them. I am a user of OCR programs for Western languages and recognize their importance to anybody needing to move printed text to computer media.

My observation to Jan Hoogland is that his description of the product he reports on — with much appreciation from all of us — does not match high enthusiasm for the product. According to Jan Hoogland, it takes several steps of preparation of the text before it can be recognized by the OCR program, and the output of the program is about 8 pages per hour. Even after ....assuming that those pages are average printed book pages, I find the OCR program to be very slow and cumbersome. I managed better than 20 pages per hour scanning the text of Yusuf Ali's "The Holy Qur'an" including the foot notes. There was no need on my part to make any preparation of the pages which consist of three main parts each: English text, footnotes, and Arabic text. Many a time I scanned 60 or more pages at a time and unless I asked to defer recognition, the program would recognize all what I have scanned as soon as I stop scanning. The 20 pages per hour, I mentioned earlier from start to finish including the final correction and cleaning of the final text. At the end of all this I have the opportunity to save my work in any one of several formats including plain text.

From what I have read about Arabic OCR packages, there is nothing that may come close to what I have described, thus my question to the members of this list. Why should Arabic users of computers allow themselves to accept inferior products at exorbitant prices? Let us not blame it on the vendors. It is our responsibility to demand better products and refuse to use inferior ones. I would love to have a good and reasonably priced Arabic OCR program, but at the rate things are progressing in Arabic computing, I have a long wait ahead of me.

Sabri Kawash

I have been the PI of ARPA Arabic OCR project which started in July 1992 and ended in July 1995. The Arabic OCR research and the obtained results by other researchers by that time were poor. Most of them had used a very small database which had no real-world importance. My goal was to implement a system using neural network technology that can recognize average to good quality machine-printed Arabic text without retraining (the ultimate goal is a character accuracy of 98%). We had access to a database of 350 documents from magazines, newspapers, and books. The data was later enhanced by inclusion of 15 Windows fonts as well. The first Beta version of the system was made available in July 1995. ...As a researcher & a technology developer who has spent some time to develop algorithms for OCR of Arabic text, I have a few comments to make here. From a user point of view, your comments are true. The Arabic OCR systems are expensive and do not compare in performance to Latin OCR systems. But is this a fair comparison? — You have to consider that Latin OCR systems have been the results of hundreds of man years of research and development in at least last 30 years and millions of dollars in investment. Arabic OCR is in its infancy and there is not a whole lot of money being spent on it, — Latin text is much less complex to recognize compared to Arabic text or Chinese for instance. The complexity of Arabic/Farsi OCR is much more than Latin text OCR. You cannot compare apples and oranges. Also when the quality of Latin text documents decreases, the performance of these systems drops sharply as well, — the extent that a technology can be developed and enhanced is directly related to the investment in that technology. If there is money, it will absorb great talents to de-

velop and enhance the technology. To invest in any technology, one needs a reasonable rate of return on his/her investment. I do not know if the Arabic OCR market (low end users) is big enough to justify the risk for anyone. As a result, prices are high and performance is lower than comparable Latin OCR systems (add to that the non-existence of Software Copy Right Law). Still with those high prices, I am sure the vendors are not making money on it, — I have developed an OCR system for Arabic which is based on neural networks and does not require training — not sensitive to size or font- as long as the text is from Naskh family of fonts or something in that neighborhood (for other styles, a retraining is required).This is a superior technology compared to other Arabic OCR systems and is in Beta test now. Will it be available for end users any time soon given the market size? I guess not. Even Microsoft is not yet making any money on their Arabic products (correct me if I am wrong). Just imagine anybody who runs Arabic Windows is probably a potential Word for Arabic customer. When it gets to Arabic OCR, there is a small portion of this population that are potential OCR customers and currently it is not enough to make developers such as myself and technology investors excited. The high-end market, may be another story. In conclusion, I think if there was a Perfect Arabic OCR system, it would not make enough sales (as a low-end product) to support and maintain itself. So you are right that there is a long wait.... and this was our side of the story!

Khosrow M. Hassibi

Dear Dr. Hassibi,

Thank you very much for your reply and clarifications, but let me assure you that aside from your involvement in OCR software development, that all the points you made are well known to me and that is why I am so concerned about the cost and quality of software available for computer users who use the Arabic character set and its variations. The market is not insignificant, but not mature yet. Any investor in that market has to look for the future and not at the present. I don't know if Microsoft is making or losing money in that market, but they have managed to dominate it and set the standards for it while the users of those languages that use the Arabic character set are passively watching. Consider the mess we have with the serious standards created by serious vendors including Microsoft, and consider the disregard Microsoft exhibited when it disregarded the standards after they were set by introducing yet another standard for Arabic Windows. What objection was there from the users? None.

My criticism of the quality and the cost of software for the Arabic computer user — by Arabic I don't mean the language but the character set — was not aimed at the software developer or the vendor as much as it was aimed at the user.

My friends, technology is very expensive and cannot be developed at the expense of the end user especially during the development stages. Among the users of the Arabic characters there are many wealth countries, some of the wealthiest in the world. Certainly some of the governments of those countries can afford to make an investment in the future of their people and spend some money to develop some software that is commercially not feasible but important and necessary such as good OCR programs. Among the readers of this form there are people who are associated with some institutions in countries were millions of dollars are spend daily on less important things that the development of good software, and I am sure that some of those people are in a position to make such suggestions to their institutions and governments. But first a realization has to be made. If people do not demand quality for a fair price, they will continue to receive mediocre products at the exorbitant prices and that not everything can be bought at the open market. Something have to be home-grown and developed at home to meet the local needs.

Sabri Kawash

I think you made the point very clear. I 100% agree with you about what you have mentioned above. Guess some of these governments must invest in some of these technologies the same way that U.S. government has invested in areas of technology that did not yet have a mature market. Specially if these technologies are directly related to their culture and language.
I hope your message will be heard.... Thank You for your thoughtful comments,

Khosrow Hassibi

The logical answer to this OCR debate is competition. If I were to develop a good Arabic OCR program, I would sell it for like $200. The other OCR developers who are selling theirs for $2000 will be forced, whether they like or not, to lower the price of theirs. Look around you and see how software and hardware prices have dropped dramatically because of fierce competition. Companies think of quality and customer service. If you're thinking of return on your investment, would you rather sell a $2000 product to 1000 users or a $200 product 10.000 users? When you're planning to develop an upgrade, which market would you choose, the 1000 users or 10.000 users? Once more good players join the game, the game will be more exciting!

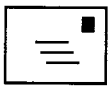
Abdel-Hadi

Well. The point I was trying to make was that the market is small TODAY and it will be hard to attract new companies to COMPETE in it. TOMORROW things may be different. With such a small market, you will not get COMPETITION in near future.

Khosrow Hassibi

I don't think the market is that small that you can't justify the development of an OCR program. If you're reasonably priced, people will buy the product and you'll have a market. Some companies take the risk of putting out a new product where they don't know how successful it will be.

Abdel-Hadi

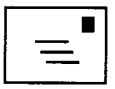


I do understand the remarks of both colleagues, but fully disagree on the point that OCR-software is of poor quality. Al-Qari' al-Ali performs very good and certainly meets my demands (although not all of them). And I do not think my demands are too low, what I need is machine-readable text that can be used for (manual) dictionary making. As I promised, I will keep you informed, and try to get some ftp-experience in order to make some files available in order for you all to judge yourselves.

Jan Hoogland

My research has focused on Arabic spell-checking, which I believe is an essential supporting component of successful Arabic OCR. However, effective Arabic spell-checking requires syntactic as well as morphological analysis, and I am unaware of any OCR system that incorporates both morphological and syntactical analysis. ...I have been impressed by Jan Hoogland's report for high OCR accuracy using Al-Qari' al-Ali (well, maybe not 99% accurate, but darn good), but the main problem for me is that I cannot see adding yet another OS (Sakhr) to my system. In fact, doesn't use of Al-Qari' al-Ali imply a "dedicated" Sakhr system? (this isn't a rhetorical question–I'd really like to know). I'd like to see a reasonably-priced (less than $1500) Arabic OCR that runs in Arabic Windows, and has the accuracy of Al-Qari' al-Ali (or better). Apparently, this requires not only OCR but some intelligent spell-checking as well. By "intelligent" spell-checking I mean contextual analysis of syntax: there are too many Arabic words that look fine (i.e. accurately spelled) in isolation, but in context you can tell right away what's wrong. Will such a product be available soon?

Tim Buckwalter



Sakhr plans to make all their products available for MS Windows and I think Al-Qari' al-Ali was on top of their list and pretty sure it's available now.

Contact Digitek in the US. I guess you mean a grammar checker. Al-Qari' al-Ali supposedly had a grammar checker included.

Abdel-Hadi

I think Al-Qari' al-Ali is already available for MS Arabic Windows...About using another system: I use three different Windows-versions: 1) Dutch Windows 2) Arabic Windows (for workgroups, etc.) 3) Arabized English windows, Nawafidh, Sakhr for the Al-Qari' al-Ali (and no Windows 95 yet, why should I?). This all works OK, as long as every Windows is in a different subdir. The only thing is that my Dutch windows shows some garbage as group- and item-titles for some programs. (But I think this can be avoided by making sure that config.sys and autoexec.bat do not make any mention of Windows, or Windows sub-directories etc.). Of course you need some extra space on your HD, but it all works very well.

Al-Qari' al-Ali for Nawafidh certainly has no grammar checker. Just a spell-checker, as I earlier described: relation between the picture-file and the text-file. Any non-existing word (according to the program's dictionary) is shown in both. But: for example the word ***mas'ūl****, as written in the correct way (i. e.* ***hamzah*** *on* ***wāw****!), is always indicated as mis-spelled. To all our Egyptian friends: your spelling (hamzah on* ***yā'*** *in* ***mas'ul****) is wrong.*

Jan Hoogland

I have constructed a web page that includes my Ph.D. dissertation on Arabic OCR and some of my papers in postscript format. The page also includes a bibliography of Arabic OCR in three different formats. The bibliography appeared in a paper titled "Survey and bibliography of Arabic text recognition" published in Siganl Processing, January 1995. You will also find a database of Arabic document images and their ground truth text, which were used in some of my experiments. You are welcome to view and download those items, but if you use them, please, cite the reference. The address of the page is http://george.ee.washington.edu/~badr

Badr Al-Badr

RUSSIAN ACADEMY OF SCIENCES
THE INSTITUTE OF ORIENTAL STUDIES
ST. PETERSBURG BRANCH

Manuscripta Orientalia

International Journal for Oriental Manuscript Research

Vol. 1 No. 3 December 1995

THESA
ST. PETERSBURG—HELSINKI

CONTENTS

Color plates: *Muraqqa`*. Album of the Indian and Persian Miniatures of the 16—18th Centuries and the Models of the Persian Calligraphy of the Same Period (see p. 63—67).

Front cover:

Fol. 17a. Portrait of a Man by Riḍā-yi ʿAbbāsī, 11.8×8.2 cm.

Back cover:

Plate 1. Fol. 16a. Portrait of Tīmūr Khān Turkmān by Ṣādiqī beg Afshār, 19.3×11.6 cm.
Plate 2. Fol. 36a. The Darvishes Picnic in the Mountains. Probably Isfahan school, 25.5×14.5 cm.
Plate 3. Fol. 6a. The Shaykh and the Harlot by Muḥammad Yūsuf Muṣavvir, 18.2×11.3 cm.
Plate 4. Fol. 1a. Portrait of Mīrzā Jalālā by ʿAlī Qulī beg Jabbādār, 16.0×9.1 cm.