Word Retrieval System for Ancient Arabic Manuscripts

Somaya Al-Maadeed, Fatima Issawi
Computer Science and Engineering Department
College of Engineering, Qatar University
Doha, Qatar
{s_alali, issawif}@qu.edu.qa

Ahmed Bouridan
Department of Computer Science and Digital Technologies
Northumbria University
Newcastle upon Tyne, UK
ahmed.bouridane@northumbria.ac.uk

***Abstract* — Many museums and libraries that have archives containing historical Arabic manuscripts are beginning to digitize their collections to enable researchers and the general public access to them. Most of these collections are available in digital format with no annotation, limiting researchers to fully utilize them efficiently. This paper reports on a system we developed for word retrieval in Historical Arabic Handwritten manuscripts. Unlike other systems, this system does not use Optical Character Recognition or word spotting techniques. The system performs three different steps: segmentation, annotation, and word search. We designed a new interface to read and query Historical Arabic Handwritten manuscripts using text queries. The queried word is then highlighted in the page image of the manuscript.**

Keywords — annotation; Arabic manuscripts; digitization; text analysis; libraries

I. INTRODUCTION

Thousands of historical Arabic manuscripts are shelved in various museums and libraries across the world that, with time, are deteriorating. Effort is being made to preserve this documented heritage in a digital format. Utilizing these manuscripts requires technologies that do more than store images of the document content. An efficient means of accessing this information requires new methods of transcribing, indexing and searching these artifacts, so that scholars can extract valuable information.

Historical Arabic Handwritten (HAH) manuscripts present a difficult problem for researchers. Arabic script is cursive and written horizontally from right to left. The shapes of letters change depending on their location in the word. Arabic words consist of a number of connected components, called Pieces of Arabic Words (PAWs). In addition, in Arabic handwritten documents, there is a lack of clear boundaries between lines, words, and between the PAWs, making word segmentation a challenging task.

Various methods have been applied to handwritten text for the preprocessing, segmentation, indexing, and word spotting to extract this valuable information. A detailed survey on Document image analysis (DIA) can be found in [1, 2, and 3].

In this project, we designed and implemented a novel segmentation, annotation, and word searching system. This paper discusses the word searching portion of our work. In that work, we present an algorithm for automatic segmentation of HAH manuscripts. The segmented page is then manually annotated to correct mistakes in segmentation. During the correction phase information about the image is extracted and stored in a database. This extracted information is then indexed and the users can use our search interface to easily find words in any ancient manuscripts that have been added to the system. To the best of our knowledge, this is the first word search system for HAH manuscripts that use text queries to highlight the search terms in the manuscript image.

This paper is organized as follows: Section II presents a literature review related to this study. Section III presents and discusses a new system for reading and searching Arabic manuscripts. An attempt to address the issue of converting handwritten documents into online searchable data is outlined. Ann overview of the initial experimental studies is presented in Section IV. Finally, we conclude our work in Section V.

II. LITERATURE REVIEW

The following is a review of the state-of-the-art methods and systems related to the study. The first subsection reviews the latest methods for retrieving words in Arabic manuscripts. Subsection B describes the available systems that provide online access to ancient manuscripts.

A. Word Retrieval for Arabic Manuscripts

There has been a lot of effort in recent years to develop word spotting systems. Word spotting systems allow the researcher to search for keywords in handwritten texts.

Word spotting methods can be classified into segmentation-based techniques, an analytical approach, or segmentation-free techniques, a holistic approach. In the analytical approach, the effort is focused on solving the word segmentation problem. In the holistic approach, the query word image is matched to the corresponding word images in the document, or template matching. All of the attempts have encountered problems dealing with handwritten Arabic. Historical documents pose a larger problem, because they are often of poor quality, with stained paper, faded ink, ink bleed, or dirt marks hindering accuracy.

Wshah et al. [4] proposed a script independent segmentation-free word spotting system based on HMMs. Not using character or word segmentation helped increase the accuracy with Arabic scripts. They used a query-by-string word spotting technique to create keyword models. Given a line image, the algorithm

978-1-5386-2756-3/17/$31.00 ©2017 IEEE

detects candidate keywords using a sliding window. Of the three languages, they tested, English, Arabic, and Devanagari, the lowest results were obtained by the Arabic scripts.

Leydier et al. [5] introduced a language independent, segmentation-free word retrieval system that allows the indexation of ancient manuscripts. Methods such as indexation are used to describe their word retrieval engine for ancient manuscripts in any language and alphabet. Other word-spotting approaches are also tested. They faced difficulty with words from the same root. They also had a low precision rate with Arabic, lower than other language databases tested.

Khayyat et al. [6], propose two verification models for Arabic word spotting systems, both using a holistic classifier to verify spotted words. They first integrate partial segmentation with a hierarchical classifier to spot words. Then two models are used for verification: first a word matching model, which passes the spotted words to a word classifier trained on lexicon words for verification. The second is a score evaluation model, which uses the scores given by the word spotting system and the holistic classifier. This system had a precision rate of 77.7%.

Khayyat et al. [7, 8], the main technique described to retrieve information from scanned handwritten documents is a word-spotting system developed to search for words within digitized documents. It is based on a hierarchical classifier that consists of a set of classifiers each trained on a different part of the input pattern. The classifier integrates a partial segmentation of lexicon words into PAWs. The components are divided into major and minor connected components. It then reconstructs and spots or rejects words using language models. Those models are also used to provide contextual information. They had a 95.4% segmentation precision rate and an 84.5% word spotting precision rate.

Aoudi et al. [9] present a word spotting approach that is a segmentation-free learning based approach using a Generalized Hough Transform (GHT) technique. It detects words by finding the model's position in the image. Their proposed system has two phases. The training phase is run once for every query word, extracting the image features from the word images using GHT and storing them. In the query resolution phase, a text query word is looked up to extract PAW models, scans the document images, and identifies all possible locations of the query word. They reached an average of 94% of well spotted words.

Makhfi et al. [10] has a lot of similarities with our research project. They propose a search engine to digitize and create a library of ancient Arabic manuscript. Makhfi et al. stresses the importance of indexing and diffusing the contents of old Arabic manuscripts either manually or automatically. The document coding is validated by XML schemas to provide format checking, type and semantics of data in XML files. Based on these metadata and annotations, the search engine provides the process of identifying, collecting and registering information. Mainly, the search engine is based on metadata and XML annotations that allow handwritten transcribed documents and indexed images to be searched in the database corresponding to users' queries.

Al-Khatib et al. [11] propose a digital library framework to fully utilize handwritten Arabic manuscripts. Their system allows the user to browse or search desired information using their interface. They use content-based retrieval from historical manuscripts that are supported by indexing systems. They start by preprocessing document images to enhance their quality, using feature extraction of 'user identified' words. They also utilize relevant feedback from the user.

El-Makhfi et al. [12] have proposed a system in their paper that is similar to our search engine. They present a new indexing system of Arabic manuscripts, to transcribe and annotate images of manuscripts according to metadata. Their system allows user to create works of Arabic manuscripts in a TEI XML format. They also supply the database of the open source platform System Documentary XML (SDX), which allows users to search, browse and view documents in a web format.

B. Search Engines for Ancient Manuscripts

There are number of available search engines for ancient manuscripts. This section lists some important resources.

OCLC WorldCat is a comprehensive database of library materials from around the world [13]. It provides access to archival and manuscript collections, mixed materials (i.e. collections that contain two or more different types of bibliographic materials), and bibliographic materials that have been described using archival descriptive rules. It also allows libraries to share their digital collections with online researchers around to world.

Many libraries provide online access to view ancient manuscripts [14, 15, and 16]. [14] provides access to search the text of the Nag Hammadi Library's translation of the Nag Hammadi Scriptures. [15] provides access to the Islamic Manuscripts collection at the University of Cambridge. They provide access to view the digital image of the book, as well as, search functionality on metadata and some content. [16] is a list of libraries using the Greenstone Digital Library software. The Armenian Rare Books collection provides access to view the digital image of the books, and provides search on metadata.

Still, the above engines do not have the ability to search for and highlight words in these manuscripts. Our application contains different manuscripts, and allows users to search on metadata, as well as for a word in all or in a specific manuscript. Our application gives the user the ability to upload new manuscript to the database and make it available for all users.

Some studies used a handwriting recognition approach to word retrieval in handwritten documents [17, 18], which search for words in the handwritten manuscript. [19-24] other studies on Arabic handwritings used different set of features for writer identifications, age recognition, and handedness recognition.

These techniques search for a word in certain line or page of a scanned manuscript. Our application will provide the ability to query for words in all available manuscripts, and highlight the query term location in all the manuscripts and documents processed. It displays the results of the query to the user who can then view the document in the document viewer. As demonstrated in the following section, the website interface is simple, helping the users find what they look for easily.

III. WORD RETRIEVAL SYSTEM

This section presents a new approach to historical text analysis and retrieving, as an attempt to help address the issue of converting historical information available on handwritten documents into electronic data, enabling it to be later searchable online. The goal is to allow users have access to historical old manuscripts without damaging the original materials.

A. System Overview

One of the traditional and common cases for word spotting is using character or shape recognition to search for words in handwritten document. Most of the available systems are font sensitive and does not suit old manuscripts. In this paper, we model a system to be used for searching and reading historical documents. As shown in Figure 1, the system starts with segmenting the document into words using our method presented in [24]. Each word in the image is annotated with the corresponding text and coordinate location. Then, for searching, we designed a database to store the segmented words' locations connected to the document and its metadata. Figure 1 represents the system process of word spotting in historical documents. In this approach, we address the process of word retrieval as an inverted process of handwriting recognition.

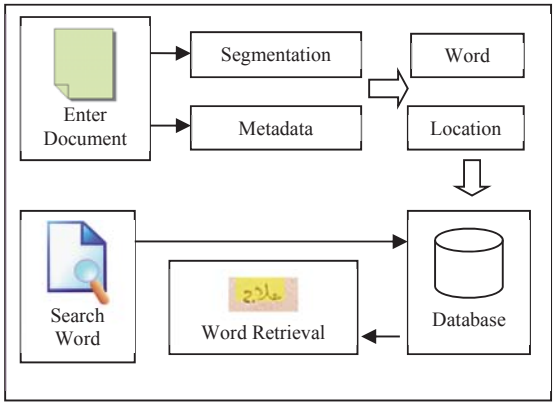

Fig 1. Overview of Word Retrieval System

B. Database structure

The database was built using SQL Server. As shown in Figure 2, the main tables in the database are: Documents, Metadatas, Pages, and Words. The main role of these tables is to make the page itself searchable. In order to make the image searchable, we needed to specify the coordinates of the word by x and y axis. This segmentation and annotation procedure is described in [24]. To test the project's functionality we used a historical Arabic handwritten book called *The Treatment of Dangerous Diseases Appearing Superficially on the Body* written by Muhammad ibn al-Hasan ibn al-Kattani. The book was written in the 11th century. Section IV describes this dataset in more detail.

C. Interface design and functionality

The system's main search functionalities are searching for a book, as well as searching for a word in the books or manuscripts. What is new about this system is that we provide users the ability to search for a word and the interface will show/highlight the corresponding word image in the book. Moreover, the user will be able to search for the desired book using searching the metadata such as keyword, title or author name. Furthermore, as shown in Figure 3, the user can click the

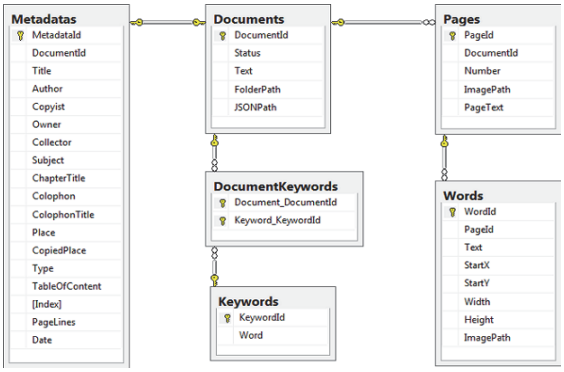

Fig 2. Database Structure of Word Retrieval System

Advanced Search button to narrow his/her search to get more accurate result. When the user views a book, the book is shown either as a gallery or as a continuous single page steam. The user can zoom in and zoom out of the pages, request to print pages from the book and he/she can search for a keyword in the book. Search results will be displayed and when the user selects the desirable word, the page is displayed with the requested text highlighted, as shown in Figure 3. Furthermore, the user can go back to the home page to search for another book. Additionally, the user can request to print the pages that he/she wants by entering the pages numbers and the number of copies that he/she wants to print. Then, the request will be submitted to the website administration. Users can also download the selected document pages as well.

IV. EXPERIMENTAL STUDIES

A. Database Used

Ibn Sina database [25] is used in the testing phase. The database contains old manuscripts that written by Sayf al-Din Abu al-Hasan Ali ibn Abi Ali ibn Muhammad al-Amidi (d. 1243A.D). These manuscripts were provided by the Institute of Islamic Studies (IIS) at McGill University in Montreal, Canada. The title of the manuscript is Kitab Kashf Al-Tamwihat Sharh Al-Tanbihat. In the period between the late twelfth century and the first half of the fourteenth century, much attention was paid

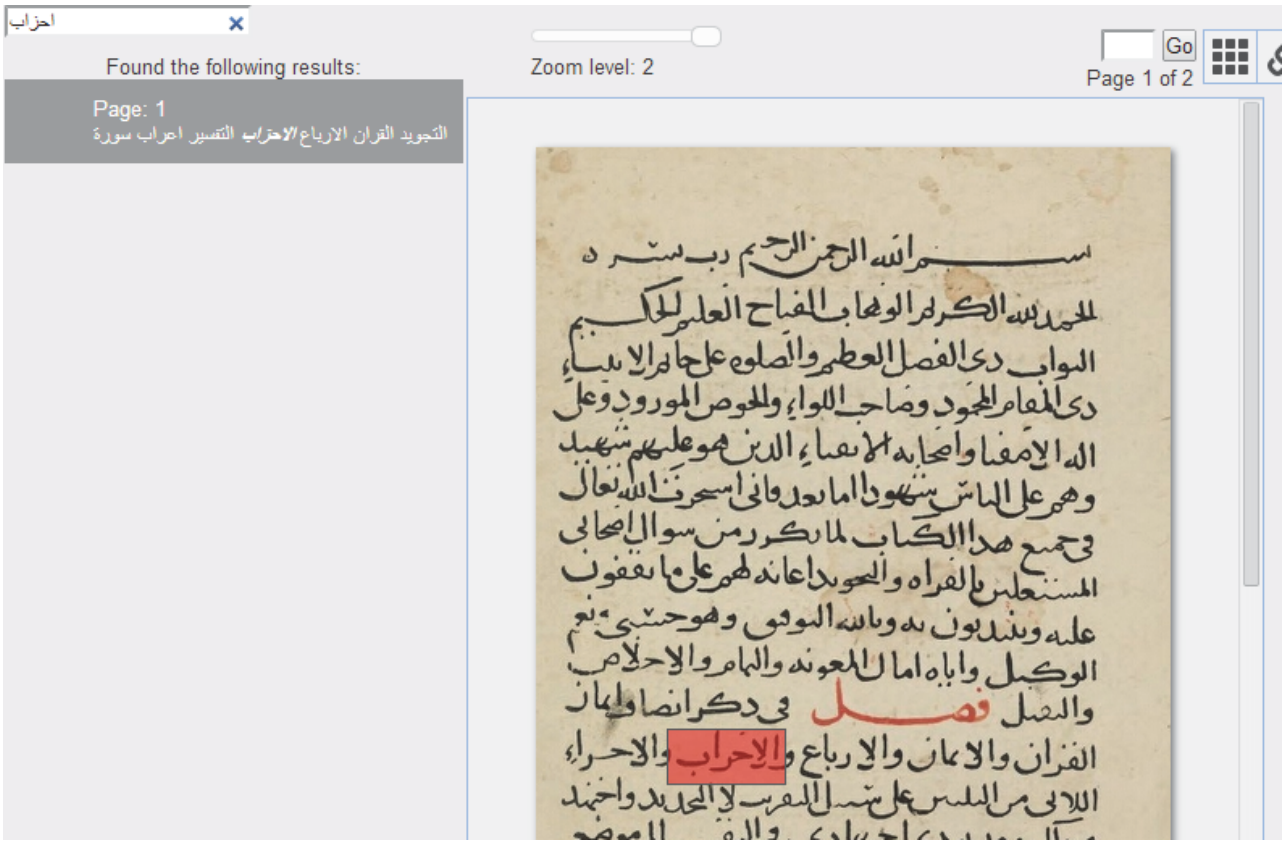

Fig 3. Word Search Interface

to this work, when more than a dozen complete observations were composed. The selected dataset form 51 folios and corresponds to 20722 extracted shapes (connected components) of Arabic script extracted from the historical manuscript. These shapes' labels were verified by the experts at McGill ISI. The Ibn Sina database was created to provide training data for optical shape recognition (OSR) and it is used in learning challenges.

B. Technical Details

In order to develop this project we use Microsoft Visual Studio 2012 and Microsoft SQL Server 2012. We developed the website using ASP.NET. The document viewer we chose is Diva.js, a Document Image Viewer with AJAX, developed by the Distributed Digital Music Archives and Libraries Lab (DDMAL) at McGill University in Montreal, Canada [25]. This open source document viewer is written in jQuery. The Diva.js plug in requires the IIPServer Image Server to serve the document images. The code was integrated and customized to meet this project's requirements. To index the annotated data in the database, we used Solr, an open source enterprise search platform by Apache. Solr's features used in this project are the full-text search, hit highlighting, indexing of the database data.

V. CONCLUSION

There are number of technical challenges that required attention during the development of this project. The main challenge is still the segmentation of HAH manuscripts to extract the searchable words. Once provided with a method to manually correct and annotate errors in the segmentation of the manuscript, the rest of the focus discussed in this paper is one that has not been developed previously for HAH manuscripts: providing a framework for the word search engine, to index, store, and provide users with searching and highlighting capability in the document image.

We considered the need for converting the words available in handwritten documents into electronic data with the goal of enabling it to become searchable online. A system prototype applying the proposed and described approach is being developed and experimentally tested, to fully demonstrate the capabilities of the proposed system on real-world data. An overview of the initial experimental studies is presented. We expect the proposed word retrieval system to take the search in HAH manuscripts to a new level.

ACKNOWLEDGMENT

This paper was made possible by a QUCP award QUCP-CENG-CSE-15-16-1] from the Qatar University. The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] A. Al-Shatnawi, F. Al-Zawaideh, S. Al-Salaimeh, and K. Omar. Offline Arabic Text Recognition – An Overview, World of Computer Science and Information Technology Journal (WCSIT) Vol. 1, No. 5, 2001, pp. 184-192.

[2] G. Nagy. Twenty years of document image analysis in PAMI, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 22, Issue 1, January 2000, pp. 38-62.

[3] J. Al-Khateeb, J. Ren, J. Jiang, and J. Al-Muhtaseb. Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking, Pattern Recognition Letters, 32, 2011.

[4] S. Wshah, G. Kumar, and V. Govindaraju. Statistical script independent word spotting in offline handwritten documents, Pattern Recognition, 47, 2014, pp. 1039-1050.

[5] Y. Leydier, F. L. Bourgeouis, and H. Emptoz. Text search for medieval manuscript images. Pattern Recognition, 40(12), 2007, pp. 3552-3567.

[6] M. Khayyat, L. Lam, and C.Y. Suen. Verification of Hierarchical Classifier Results for Handwritten Arabic Word Spotting, 2013 12th International Conference on Document Analysis and Recognition, pp. 572-576.

[7] M. Khayyat, L. Lam, and C.Y. Suen. Learning-based word spotting system for Arabic handwritten documents, Pattern Recognition, 47, 2014, pp. 1021-1030.

[8] M. Khayyat, L. Lam, and C.Y. Suen. Arabic Handwritten Word Spotting Using Language Models, 2012 International Conferenct on Frontiers in Handwriting Recognition, pp. 43-48.

[9] N. Aouadi and A. Kacem. OCR-Independent and Segmentation-Free Word Spotting in Handwritten Arabic Archive Documents, 2013 International Conference on Electrical Engineering and Software Applications (ICEESA), pp. 1-6.

[10] N. Makhfi, O. Bannay, and R. Benslimane. Search engine of ancient Arabic manuscripts based on metadata and XML annotations, Information Science and Technology Conference (CIST), May 2011, pp. 1-10.

[11] W. Al-Khatib, S. Shahab, and A. Mahmoud. Digital Library Framework for Arabic Manuscripts, International Conference, May 2007, pp.458-465.

[12] N. El-Makhfi, O. El-Bannay, and R. Benslimane. System of indexing, annotation and search in the old Arabic manuscripts, Information Science and Technology Conference (CIST), May 2011, p.9.

[13] K. Raseroka,and S. Mutula, (2012). Retracing the impact of Information Communication Technology on academic libraries in sub Saharan Africa: Case study of the University of Botswana Library. In. Ravindra N. Sharma (ed). libraries in the Early 21st Century: An International Perspective (Vol II). Berlin: De Gruyster Saur/IFLA, 2012. pp 129-144.

[14] Nag Hammadi Library (The Nag Hammadi Scriptures) [Online] Available: http://www.gnosis.org/naghamm/nhl.html

[15] Cambridge Digital Library [Online] Available: http://cudl.lib.cam.ac.uk/collections/islamic.

[16] Greenstone Digital Library software [Online] Available: http://www.greenstone.org/examples.

[17] K. Khurram, F. Claudie, and V. Nicole, Word spotting in historical printed documents using shape and sequence comparisons, University Paris Descartes: France, vol. 45(issue 7), pp. 2598-2609, July 2012

[18] F. Ethem, and D. Pına, A line-based representation for matching words in historical manuscripts, Bilkent University: Turkey, vol. 32(Issue8), pp.1126-1138 , June 2011.

[19] S Al-Maadeed, F Ferjani, S Elloumi, A Jaoua, A novel approach for handedness detection from off-line handwriting using fuzzy conceptual reduction, EURASIP Journal on Image and Video Processing 2016 (1), 1.

[20] E. Khalifa, S. Almaadeed, M. Tahir, A. Bouridane, A. Jamshid, Off-line Writer Identification Using an Ensemble of Grapheme Codebook FeaturesPattern Recognition Letters, published online: 17-APR-2015 Full bibliographic details: Pattern Recognition Letters (2015), pp. 18-25

[21] S. Almaadeed, A. Hassaine, A Bouridane, M. Tahir, Novel geometric features for off-line writer identification, Pattern Analysis and Applications, 27 Dec 2014.

[22] S. Al-Maadeed, A. Hassaine, Automatic prediction of age, gender, and nationality in offline handwriting, EURASIP Journal on Image and Video Processing, P 1-10, vol. 2014, no. 1, January, 2014.

[23] S. Al-Maadeed, Text-Dependent Writer Identification for Arabic Handwriting, Journal of Electrical and Computer Engineering, Hindawi, doi:10.1155/2012/794106, vol. 2012, March 2012.

[24] A. Hassaine. A robust method for line and word segmentation in handwritten text. Qatar Foundation Annual Research Forum Proceedings: vol. 2013, ICTP 057.

[25] R. Farrahi Moghaddam, M. Cheriet, M. Adankon, K. Filonenko, and R. Wisnovsky, IBN SINA: A database for research on processing and understanding of Arabic manuscripts images"\, Proceedings of DAS'10, June 9-11, 2010, Boston, MA, USA