WAHD: A database for Writer Identification of Arabic Historical Documents

Alaa Abdelhaleem*, Ahmed Droby*, Abedelkader Asi, Majeed Kassis, Reem Al Asam, Jihad El-sanaa

***Abstract*—A comprehensive Arabic handwritten text database is an important resource for Arabic handwritten text recognition research. It is essential for training text recognition algorithms and vital for evaluating the performance of these algorithms. In this paper, we present a database that includes manuscripts from the Islamic heritage project (IHP), consisting of 333 historical manuscripts written by 302 different writers, 23 from them are known. The database contains 54 manuscripts, whose writers are known, from 13 sources. Among these known writers, 11 have written multiple manuscripts. The total number of pages in the entire database is 36,969. Each manuscript in the database accompanied with metadata that include various properties of the manuscript, such as title, creator, subject, language, copyist name, etc. To enrich our database we added twenty historical books scanned from the National Library(NLJ), in Jerusalem. The books have different number of pages and different writing styles. In addition, we present a number of experimental results on the database using two classifiers, The GMMS System and The OBI/SIFT System. The database is made freely available to researchers worldwide for research in various handwritten related problems such as text recognition, writer identification, verification, forms analysis, pre-processing and segmentation.**

I. Introduction

Historical documents images have been attracting the interest of scholars from various disciplines, each from his/her own viewpoint. Transforming these documents into digital images has invited the interest of computer science researchers, and may algorithms and systems were developed to process historical document. Since the research on digital document images is usually more toward an applied one, the availability of the real and diverse databases is crucial, especially for handwritten historical documents.

The Arabic heritage is a among the richest in terms of quality and quantity, at least until the $17th$ century. The vast majority of this heritage is textual. The original manuscripts that was written before the $17th$ century are handwritten.

Obtaining a large and diverse database that represents the variety of handwriting styles and contains the most important classes in the target language is crucial for the development of efficient algorithm for processing historical-handwritten document images. Existing databases doesn't have a large enough samples to enable the use of a more advanced recognition and classification techniques, e.g. typical neural network need a large training data to work well. Thus the lack of a large database limit the development of the research in this area.

In this paper we present a large and comprehensive Arabic handwritten text database, which consist of two subsets IHP and NLJ. In addition, we present experimental results on IHP subset using two classifiers.

This paper is organized as follows: Section II describes related works toward developing different Arabic historical handwriting databases. The WAHD database and its specifications is presented in Section III. In Section III-A, we present the data collection stage. Some statistics about the database is presented in section III-B. In Section IV, the experimental results on Arabic text recognition using part of the data are presented. Finally, we present our conclusions and direction for future work.

II. Related Work

In the last twenty years, a lot of diverse and huge databases has been created for handwritten Latin-scripts [1] [2] [3]. However, there has not been much effort toward developing comprehensive databases for Arabic handwriting recognition, and most of them made in laboratory conditions and do not represent the changing handwritten over the history and the variations in the handwriting styles of the writer over a period of time. The existing databases, can be divided into two categories: databases that target text recognition and those that target writer identification.

IFN/ENIT database was developed in 2002, by the Institute of Communications Technology (IFN) at Technical University Braunschweig in Germany and the The National School of Engineers of Tunis (ENIT). It consists of 26,549 images of Tunisian town/village names written by 411 writers [4]. It is one of the most widely used databases. Although it has a little number of vocabulary because it contains mainly names of towns and villages of Tunisia.

The Al-Ohali *et al.* [5] database was released in 2003 by the Center for Pattern Recognition and Machine Intelligence (CENPARMI) and based on Arabic check database. It was geared toward research in recognition of Arabic handwritten checks. The database includes images for Arabic legal amounts, and Arabic sub-words (mainly used in writing legal amounts, courtesy amounts, and Indian digits). Al ISRA database [6] contains Arabic words, digits, signatures, and free form Arabic sentences gathered from 500 randomly selected students at Al Isra University in Amman, Jordan. Al ISRA database was collected by a group of researchers at the University of British Columbia. This database has the same limitation regarding Arabic text as it is made of words, digits, signatures, sentences and not normal Arabic paragraphs of text. The AHDB database, developed in 2003 by Al Maadeed [7], and includes images of words that are used to describe numbers and quantities in checks, images of the most frequent words used in Arabic writing, and images

978-1-5090-6628-5/17/$31.00 ©2017 IEEE

Database	# writers	Description
WAHD	322	333 manuscripts for 302 different writers, And 20 book for 20 writers. Over all 43,976 pages.
KHATT	1000	1000 forms, 2000 (random and fixed paragraphs) & free paragraphs
Al-Isra [6]	500	37,000 words, 10,000 digits, 2,500 signatures, 500 sentences
IFN/ENIT [4]	411	26,459 images of Tunisian city names
Alamri et al. [9]	328	46,800 digits, 13,439 numerical strings, 21,426 letters, 11,375 words, 1,640 special symbols

TABLE I
SHOWS FAMOUS DATABASES ON ARABIC HANDWRITTEN CHARACTERS, WORDS, TEXT, AND DIGITS.

of sentences used in writing legal amount on Arabic checks. The KHATT database [8] includes unconstrained handwritten Arabic text written by 1000 different writers. It was developed jointly by research groups from KFUPM, Saudi Arabia, TU-Dortmund, Germany, and TU-Braunschweig, Germany.

III. DATABASE

In this paper we present a for writer identification of Arabic historical documents (WAHD), which includes almost full historical handwritten Arabic manuscripts from different writers over various periods of time. Currently, most of the manuscripts were collected from the Islamic Heritage Project (IHP) and the National Library in Jerusalem (NLJ). The database is meant to be dynamic and more manuscripts is expected to be added to this database. It currently includes 353 manuscripts, 333 from IHP and 20 from NLJ. WAHD database is freely available to interested researchers [10].

The IHP manuscripts, which contain degradation, decorations, and margin notes, were written by 302 different writers, 23 of them are known. Eleven scribes wrote 42 manuscripts consisting of $2,313$ pages, where each scribe wrote more than one manuscript. Let us denote the set of pages written by these 11 scribes by *S-Multi*. The rest of the known writers, 12, wrote one manuscript each. These manuscripts have $2,108$ pages in total and they are denoted by *S-single*. The writers of the remaining 279 manuscripts are unknown and the set of pages in these manuscript is denoted as *S-unknown* and contains in total $32,548$ pages. The number of pages in the IHP subset is $36,969$. The NLJ manuscripts cover 6 different topics: Religion, Mathematics, Biology, Physics, Agriculture and Literature. The books were photographed using a high quality camera, namely Hasselblad H5D-60 Medium Format Digital SLR Camera from 1m distance. They are stored in an uncompressed TIFF format, where each image is roughly of size 6000×6000 pixels. Each image is roughly 100 MB of size, and due to size limitations the released database contains images of reduced size. The subset of the database are still quite large, each file is roughly 1GB. Most of the books have more than 100 pages (As shown in table IV). The books written in different centuries from 15th to 20th century. The number of pages in the NLJ subset is 7007.

The presented database have a number of advantages over existing databases. It is the largest, publicly available, database of historical Arabic documents, in terms of page count. The availability of many pages for each writer that were written over long period of time is beneficial for the development of more complex research in writer identification. For example, the existence of multiple manuscripts (S-Multi) for a scribe is vital for writer classification, and the fact that these manuscripts provide sample of handwriting over various periods of time for the same scribe could provide insight over the evolvement of handwriting of an individual over time. One could also study the different writing conditions (i.e., mental condition, time of day/month, writing tool etc.) from the variation in writing of an individual scribe. We also believe this database will contribute to the study of the development of the Arabic script and the various shapes of letters over time and across different geographic regions.

Later in the paper we present a test results from two different writer identification and classification methods, and we invite researchers to evaluate their method using the presented database.

A. Data Collection

Arabic historical manuscripts are scattered all over the world in archives, libraries, and private collections. Each individual or organization has his own policy for provide databases for research, open-to-public, or commercial use. In addition, some of the obtained manuscripts include various types of noise and they may require some pre-processing to remove noise and undesirable regions in many pages (marginal notes), which require sophisticated image segmentation.

The manuscripts from IHP include two types of noise that may affect the performance of the writing style identification and retrieval tasks: (i) the scanner surface is included in the manuscript scans, and (ii) decorations and notes appear in page margins. We eliminate the scanning background using the color differences between the scanner surface and the manuscripts pages. Colors were represented in the CIELAB color space to cluster the page pixels into two clusters. It is widely considered that this color space is perceptually uniform for small color distances [11]; a property that ensures adequate clustering despite the presence of aging noise. The result of this step is illustrated in Figure 1(2). Due to the inherent noise contained in ancient manuscript pages, e.g. decorations and notes on page margins, we apply the method suggested by Asi *et al.* [12] to detect and extract the main text region only. Historians determine that notes on page margins were added along the years by different writers. The authors suggested a technique that utilizes the unique texture and orientation imposed by the main text with respect to text in margins. Following this observation, they employ Gabor filter as it had been found to be particularly appropriate to distinguish between texture representations [13]. In the final step, they refine the gabor-based coarse segmentation using Markov Random Fields which produces a binarized version of the main-text region as appears in Figure 1(3).

Fig. 1. Pre-processing pipeline for the IHP database. (1) Original image (2) Background cropping, and (3) Main-text segmentation.

Fig. 2. two samples from NLJ database with high resolution.

Country	#	Country	#
china	3	Serbia	1
egypt	8	Syria	34
Greece	1	Turkey	18
India	55	Uzbekistan	1
Iran	4	Pakistan	1
Lebanon	1	Unknown	201
Morocco	5		
Total:			333

TABLE II
THE ORIGIN OF THE VARIOUS MANUSCRIPTS IN THE DATABASE.

Century	#	Century	#
13	10	18	47
14	10	19	59
15	9	20	22
16	32	Unknown	98
17	46	Total	333

TABLE III
MANUSCRIPTS CENTURY.

Id	# pages	Id	# pages
1	42	12	564
2	264	13	422
3	120	14	420
4	484	15	196
5	78	16	392
6	58	17	190
7	11	18	1414
8	114	19	622
9	1042	20	372
10	72	**Total**	**7007**
11	130		

TABLE IV
STATISTICS ABOUT NATIONAL LIBRARY BOOKS.

B. Statistics

1) IHP subset: The IHP manuscripts originated from 13 different regions, which are listed in Table II. The unknown label indicates unknown region of the specific manuscript. According to metadata the manuscripts were written in different centuries. Table III shows the number of manuscripts that written in the different centuries.

2) NLJ subset: Table 4 shows the currently available metadata for each book in the NLJ subset. e.g , century, subject, and number of pages.

IV. EXPERIMENTS

For our experiment we ran two writer identification systems only on IHP manuscripts [10], The GMMS System [14] and The OBI/SIFT System [12].

A. The GMMS System

To make the database valid for GMMS System [14], the authors apply a pre-processing phase. They used some heuristics, such as (A) compute the number of connected components

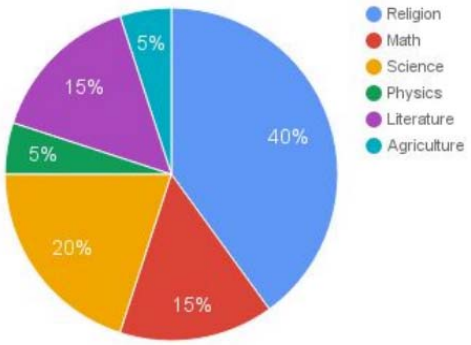

Fig. 3. Distribution of the subjects in NLJ subset.

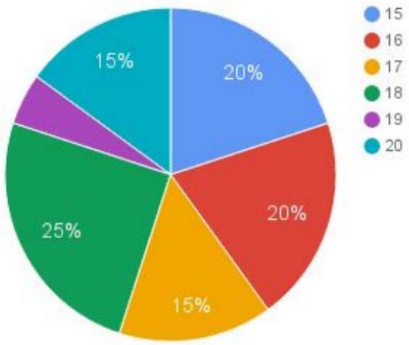

Fig. 4. Distribution of the centuries in NLJ subset.

after applying the closing-morphological operator, for each image, (B) compute the mean and the standard deviation of all the number in the connected components (C) select all images having more than a certain number of connected components (CC). Eight manuscripts that did not have a sufficient number of pages for classification were removed.

"The GMMS approach is based on [14], in which the authors propose to use GMM super-vectors to encode the features of a document. First, RootSIFT descriptors are extracted for each document. RootSIFT is a variant of SIFT in which the features are additionally normalized using the square root (Hellinger) kernel. The descriptors from the training set are used to train a Gaussian mixture model (GMM). The GMM parameters are estimated using the expectation-maximization (EM) algorithm. This GMM serves as a universal background model (UBM) from which document specific GMMs are computed. The UBM is adapted to the features of the query document by means of one Maximum-A-Posteriori (MAP) step, followed by a mixing step. In the MAP step the new statistics up to the second order are computed which are then mixed (using a relevance factor) with the parameters of the UBM to create the document specific GMM. The parameters of this newly created GMM are concatenated to form a super-vector. After a normalization step, this high dimensional super-vector is used for comparison using the cosine distance. For classification a 1-nearest-neighbor classifier is used, i.e. for each document its GMM super-vector is computed and compared to all GMM super-vectors of the training set (thus, apart from the GMM no training is involved). Note that in contrast to [14], the orientation information of the SIFT keypoints has been dropped. Furthermore, instead of a power-normalization of the GMM super-vectors, a component wise L_2 normalization is applied before the global L_2 normalization" [15].

Top_1	0.94946188	Top_6	0.98892528
Top_2	0.98268598	Top_7	0.98954922
Top_3	0.98674154	Top_8	0.99001718
Top_4	0.98783344	Top_9	0.99048513
Top_5	0.98876929	Top_{10}	0.99095303

TABLE V
WRITER IDENTICATION ACCURACY ON THE IHP DATABASE USING ROOTSIFT FEATURE.

Table V presents the result of computing the Top_k (k=1...10) of writer classification; e.g., every writer have pages in training and in testing. The number of images in the training phase is $25,030$, and in the testing phase is $6,411$. These images are classified into 294 classes.

B. The OBI/SIFT System

The employed system uses a combination of two potent feature extraction techniques. The first approach is the oriented basic image (OBI) feature extraction technique [16] which works on a textural level. Multi-scale symmetry and orientation features are computed for each pixel of the query document image by employing a Gaussian derivative filter pyramid. A histogram of these features is computed and used as the feature vector for the writer identification. The second feature extraction technique utilized key point-based descriptors (SIFT) which is presented in [17]. A feature vector is generated by computing the cosine distances between all the elements of the extracted SIFT descriptor-vectors of a query document image. The OBI features are extracted from binarized images (obtained by applying the Otsus method) and the SIFT-based features are computed on the gray-scale images. The classification is done using a special k-nearest-neighbor approach which normalizes the distances of a query document to the k-nearest neighbors in a training database, similar to [17], and then weighted the normalized distances by the performance of each feature, which are used to create a histogram. The entries of this histogram are the various writers. The index with the maximum value in this histogram corresponds to the classified writer of the query document. The other entries of the ranked list for calculating the Top_5 and Top_{10} performance consist of the results of the single features which are employed alternately. These results were obtained using $k = 3$ and $\sigma = 2.5$ as a base for the Gaussian derivative filter pyramid of the OBI features. Furthermore, χ^2-distance is utilized as distance metric for classification.

Feature	Page Level	Averaging	Voting	W-Voting
G-SIFT	0.71	0.79	0.76	0.81
HR-SIFT	0.60	0.61	0.73	0.76
HE-SIFT	0.72	0.76	0.79	0.79
OBI	0.63	0.76	0.76	0.79

TABLE VI
STATISTICS ABOUT NATIONAL LIBRARY BOOKS.

In this experiment we study the distinctiveness of the presented features and the classication schemes. Toward this end, the set $S - Multi$ is specied as the testing set to ensure that the reference set includes at least one manuscript written by the same writer of the query manuscript. We iteratively select a query Q manuscript, $Q \in S - Multi$, while the rest of the manuscripts, $(S - Multi \setminus Q) \cup S - Single \cup S - Unknown$, behave as a reference set to utilize the broader variability of the full database.

The G-SIFT feature provides the highest identification accuracy when combined with the weighted voting classication scheme. Hessian regions descriptors yield better results than Harris regions ones, an observation which is consistent with a previous comprehensive study on descriptors performance [18].

V. CONCLUSION

We have presented a large database collected from two different sources the Islamic Heritage Project and the National Library, in Jerusalem. The database consist of 353 manuscripts written by 322 scribes, part of them are known. It contains a total of 43,976 pages. Experimental results of the GMMS and OBI/SIFT systems on the IHP subset is reported. The GMMS system [14] provided better results that OBI/SIFT System [17] but it took more time; the GMMS system applies a learning phase, while the OBI/SIFT system computes the distances between the vectors.

This database freely available for the research community interested writer identification and verification. We expect more experimental results on the database to be published. To enrich the information available on the data, we are gathering more metadata that will be added to the database. In addition, we will be adding more subsets of books from the National Library with more rich metadata on each book.

ACKNOWLEDGMENT

This research was supported in part by the Lynn and William Frankel Center for Computer Sciences at Ben-Gurion University, Israel, and we'd like to thank them for their support.

REFERENCES

[1] G. Dimauro, S. Impedovo, R. Modugno, and G. Pirlo, "A new database for research on bank-check processing," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 524–528.

[2] U.-V. Marti and H. Bunke, "A full english sentence database for off-line handwriting recognition," in *Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on*. IEEE, 1999, pp. 705–708.

[3] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 5, pp. 550–554, 1994.

[4] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, H. Amiri *et al.*, "Ifn/enit-database of handwritten arabic words," in *Proc. of CIFED*, vol. 2. Citeseer, 2002, pp. 127–136.

[5] Y. Al-Ohali, M. Cheriet, and C. Suen, "Databases for recognition of handwritten arabic cheques," *Pattern Recognition*, vol. 36, no. 1, pp. 111–121, 2003.

[6] N. Kharma, M. Ahmed, and R. Ward, "A new comprehensive database of handwritten arabic words, numbers, and signatures used for ocr testing," in *Electrical and Computer Engineering, 1999 IEEE Canadian Conference on*, vol. 2. IEEE, 1999, pp. 766–768.

[7] S. Al-Ma'adeed, D. Elliman, and C. A. Higgins, "A data base for arabic handwritten text recognition research," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 485–489.

[8] S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Märgner, and H. El Abed, "Khatt: Arabic offline handwritten text database," in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*. IEEE, 2012, pp. 449–454.

[9] H. Alamri, J. Sadri, C. Y. Suen, and N. Nobile, "A novel comprehensive database for arabic off-line handwriting recognition," in *Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR*, vol. 8, 2008, pp. 664–669.

[10] vml, "dataset name," https://www.cs.bgu.ac.il/∼vml/wahad.html, 2016, [Online; accessed 2016].

[11] R. Cohen, A. Asi, K. Kedem, J. El-Sana, and I. Dinstein, "Robust text and drawing segmentation algorithm for historical documents," in *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*. ACM, 2013, pp. 110–117.

[12] F. D., A. A., M. V., E.-S. J., and F. T., "Writer identification for historical documents," in *In Proc. 22th International Conference on Pattern Recognition*, 2014.

[13] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biological cybernetics*, vol. 61, no. 2, pp. 103–113, 1989.

[14] V. Christlein, D. Bernecker, F. Hönig, A. Maier, and E. Angelopoulou, "Writer identification using gmm supervectors and exemplar-svms," *Pattern Recognition*, vol. 63, pp. 258–267, 2017.

[15] F. Slimane, S. Awaida, A. Mezghani, M. T. Parvez, S. Kanoun, S. A. Mahmoud, and V. Märgner, "Icfhr2014 competition on arabic writer identification using ahtid/mw and khatt databases," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 797–802.

[16] A. J. Newell and L. D. Griffin, "Natural image character recognition using oriented basic image features," in *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*. IEEE, 2011, pp. 191–196.

[17] D. Fecker, A. Asi, V. Märgner, J. El-Sana, and T. Fingscheidt, "Writer identification for historical arabic documents." in *ICPR*, 2014, pp. 3050–3055.

[18] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.