ORIGINAL PAPER

CrossMark

KERTAS: dataset for automatic dating of ancient Arabic manuscripts

Kalthoum Adam[1] · Asim Baig[1] · Somaya Al-Maadeed[1] · Ahmed Bouridane[2] · Sherine El-Menshawy[3]

Abstract

The age of a historical manuscript can be an invaluable source of information for paleographers and historians. The process of automatic manuscript age detection has inherent complexities, which are compounded by the lack of suitable datasets for algorithm testing. This paper presents a dataset of historical handwritten Arabic manuscripts designed specifically to test state-of-the-art authorship and age detection algorithms. Qatar National Library has been the main source of manuscripts for this dataset while the remaining manuscripts are open source. The dataset consists of over 2000 images taken from various handwritten Arabic manuscripts spanning fourteen centuries. In addition, a sparse representation-based approach for dating historical Arabic manuscript is also proposed. There is lack of existing datasets that provide reliable writing date and author identity as metadata. KERTAS is a new dataset of historical documents that can help researchers, historians and paleographers to automatically date Arabic manuscripts more accurately and efficiently.

1 Introduction

Islamic civilization contributed significantly to modern civilization; the period from the 8th to 14th century is known as the Islamic golden age of knowledge. This period marked an era in history when culture and knowledge thrived in the Middle East, Africa, Asia and parts of Europe. Arabic was the language of science and the Arab world was the center of knowledge [1]. Millions of Arabic manuscripts from that era on a wide variety of topics are scattered in different collections across the world. Many efforts have been made by numerous contributors to preserve this valuable heritage. Unfortunately, due to physical degradation of the paper and the ink, processing and studying these documents has proven to be a challenging process. Consequently, these documents are actively being digitized to preserve them. Historians and paleographers are encouraged to work with these

digitized versions of the manuscripts. These digital copies are very attractive to researchers because they allow quick and easy access to these historical manuscripts, which in turn provides a way to evaluate, analyze and research these documents without physically handling the delicate and precious works.

The publication or writing date of a historical manuscript has always been important for historians. It can help them understand the sub-textual context of the document and also help in understanding the cultural and historical references that are presented in the text. Knowing when the manuscript was written can also help researchers catalogue and categorize historical documents more accurately and efficiently. Traditionally, historians and paleographers have used invasive methods such as identifying the texture and composition of the paper or components used to make the ink to estimate the age of the document [2]. Some even try to find clues such as dates of historical events within the written content as well as the handwriting and punctuation in order to find the age of the document [3]. A few researchers have also studied ornamentation and watermarks in the documents in order to determine the age of these manuscripts [4]. As mentioned earlier, a large number of ancient manuscripts have been scanned and digitized by libraries and museums. These scanned images have enticed the pattern recognition community as a whole and image processing researchers in particular

✉ Somaya Al-Maadeed
  s_alali@qu.edu.qa

1 Computer Science and Engineering Department, College of Engineering, Qatar University, Doha, Qatar

2 Department of Computer and Information Sciences, Northumbria University Newcastle, Newcastle upon Tyne, UK

3 Department of Humanities, College of Arts and Science, Qatar University, Doha, Qatar

to try and solve the problem of document age detection using noninvasive techniques [5].

Classifying ancient documents based on writing styles is one of the techniques used to date these documents. System for paleographic Inspection (SPI) [6] is one of the earliest researches that employs writing style-based techniques for ancient documents dating. SPI uses tangent distance and statistical based algorithms to build models of all characters. Afterward, SPI uses the models to measure similarity of the letters in their dataset with the letters of the tested document. Moreover, He et al. in [7] proposed an approach where global and local support vector regression is used with writing style-based features (hinge and fraglets to estimate the date of historical documents. Alternative research on dating ancient manuscript [8], suggests using histogram of orientation of strokes as a feature descriptor to represent the image documents. The descriptor is later sent to self-organizing map clustering system to match the image with a date label. Similarly, Wahlberg et al. used a method based on shape context and stroke width transformation to create a statistical structure for dating ancient Swedish characters [9]. Whereas Howe et al. at [10] applied the Inkball models of isolated character for dating ancient Syriac characters.

While there are quite a few online libraries with datasets in various languages that possess thousands of manuscripts. However, most researchers had to develop their own datasets and find the authorship and age information for verification before they could test and verify their algorithms. A brief review on some existing online dataset is studied in Sect. 4.

The next section provides a brief history of Arabic handwriting over the centuries and its distinguishing characteristics in each period of Islamic history. The design process and description of KERTAS are provided in Sect. 3. Section 4 focuses on a comparison of KERTAS dataset with currently available digitized manuscript resources. Section 5 presents the proposed features to identify the age of historical handwritten Arabic manuscripts. Results and discussion is elaborated in Sect. 6. Then, conclusions are presented in Sect. 7.

2 History of Arabic writing style and manuscripts

While Arabic scripts existed before Islam, Arab was an oral society in that period. Only a few inscriptions were found that go back to that time. Figure 1 shows one of the pre-Islamic inscriptions found in south of Damascus, Syria, in the 6th century C.E. The inscription was written with al Jazm, one of the earliest known styles of the modern Arabic scripts [11].

Islamic calendar (Hijri) started at 622 C.E. Dates in Islamic calendar are denoted A.H. (Latin: Anno Hegirae).
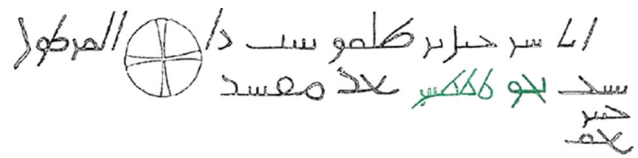

Fig. 1 Pre-Islamic Arabic Jazm inscription

Fig. 2 Early Qur'anic manuscript held by the University of Birmingham Library

In the first two Islamic centuries, most of the historical documents were Qur'anic manuscripts written in Kufic and Hijazi scripts with no signs for the short vowels and no dots to differentiate similar letters as shown in Fig. 2. After the Islamic world expanded, many non-Arab Muslims found it difficult to read the Quran and distinguish between Arabic letters. Arab grammarians adjusted the text of the Quran to avoid distortion; therefore, dots were introduced.

The second half of the second A.H. (8th C.E.) century, the world had seen a growth in the translation movement, and this period was the starting point of the new phase of manuscript history. The arrival of paper in the region contributed to the increase in material documented during that time. Paper was known in China five centuries before Islam. Arabs were introduced to paper in 751 C.E. The number of writing scripts increased. In the 3rd A.H (9th C.E.) century, six scripts were introduced. They were thuluth, naskh, tawāqīᶜ, rayḥān, muḥaqqaq and riqā [12]. The dominate scripts in the period of the 4th–7th Islamic century (10th to 13th C.E) were thuluth, naskh, tawāqīᶜ, rayḥān, muḥaqqaq, and riqā. Figure 3 shows a manuscript from the 5th A.H. (11th C.E.) century.

After 8th A.H. (14th C.E.) century, manuscript topics presented less interest in science and more in literature and poetry alongside Islamic subjects. Toward the 13th Islamic century, many of the scripts used earlier had disappeared and the following were used in a wide scale: kūfī, thuluth, naskh, ruqᶜah, faressy and tawāqīᶜ [13]. There was also a special interest in Arabic calligraphy in that period. Figure 4 shows an example of scripts used during 13th Islamic century.
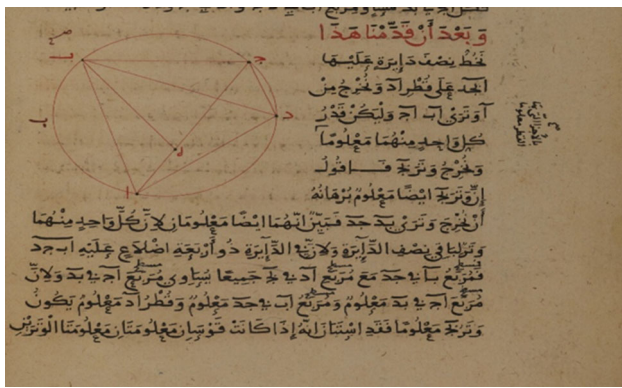
Fig. 3 Examples of mathematical problems from QNL from 5th century

Fig. 4 Al-Qaṣīdah Al-Muḥammadīyah from the University of Cambridge Digital resources, written in the 13th century

3 Dataset description

KERTAS is a term which is applied to paper in Arabic literature records, it is also called waraq [14]. Scholars of Arabic history, scripts and texts consider division on the basis of Islamic centuries to be a better fit. Therefore, manuscripts in KERTAS dataset are categorized based on the Islamic century of their publication/writing. In the development of KERTAS dataset, special care has been taken to find and gather a significant number of manuscripts from each century. We have also attempted to properly verify the writing date of each manuscript in the database. This was not a simple task especially since there are quite a few manuscripts with ambiguous or incorrect writing dates. We checked the writing dates of these manuscripts from multiple sources before adding them to the database While collecting the database, we faced another very complicated issue regarding the writing date of certain manuscripts. In the later centuries, some manuscripts were copied by hand from earlier works. In this case, what would the manuscript writing date be and who would be the author? Ideally, the date should be the writing date of the original manuscript and the author should be the author of the original manuscript. For the algorithms attempting to automatically identify the author or writing date based

Table 1 KERTAS dataset sources

Name of sources	Number of manuscripts
BRILL through QNL	57
University of Tubingen, Berlin	1
Dar al Makhtotat Sana'a Yemen	1
Institute of oriental culture, University of Tokyo	8
Princeton University Library	4
Wellcome Library	2
Yale University	4
Cambridge	
University Library	3
Islamic Awarness Web site	13
The Royal Library, National Library of Denmark	2

on the style of writing, the name of the copier and the date of copying should be selected. In this dataset, we have solved this issue by recording the name of the copier and date of copying with the manuscript. This is done in order to provide a consistent view of the data for the author and writing date recognition algorithms. Table 1 shows the sources of KERTAS dataset and the number of manuscripts that were collected from each source. KERTAS dataset not only contains a large number of images (over 2000), but they are quite nicely divided over the 14 Islamic centuries.

Table 1 also shows that Qatar National Library is the main contributor with significant contributions from Cambridge Digital Library, University of Tokyo Library, University of Tubingen, Berlin, and Princeton University Library. The storage structure of the dataset is shown in Fig. 5a. Manuscript is the main directory with a subdirectory for each century. Within each century subdirectory, there are additional directories for each manuscript. These individual manuscript directories contain images of multiple pages from these manuscripts. Figure 5b presents the content of the first manuscript in the first century. Each manuscript subdirectory also contains an XML file, which provides metadata about the manuscript including information about the century in which it was created, source, description, ID, manuscript name and writer's name if available. Figure 5c shows a sample XML file.

4 Comparison of KERTAS with existing datasets

4.1 Existing datasets

It is important to note that there are quite a few manuscript datasets available for algorithm testing and training but they
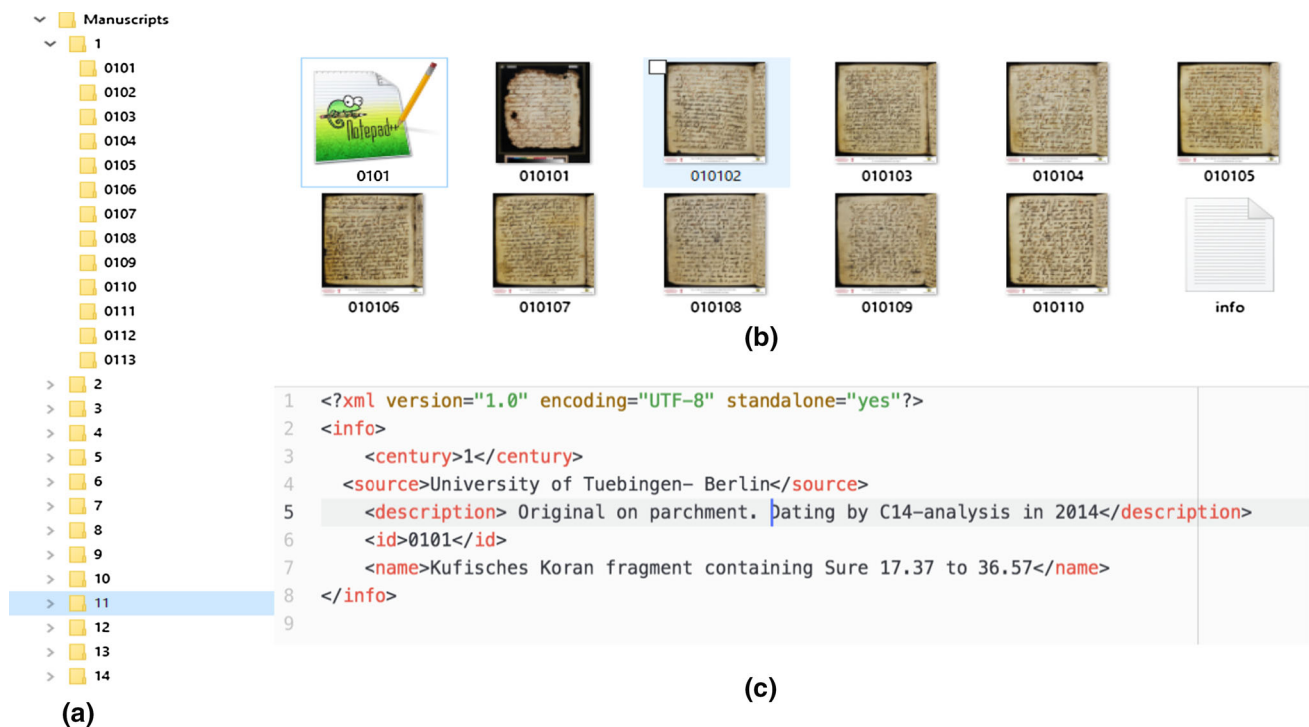
Fig. 5 **a** Directory structure for kertas database, **b** content of manuscript number 0101 in the first century and **c** xml file with manuscript metadata

all have their own limitations and advantages. The institute de recherché et d'histoire des textes (IRHT) has an online dataset of over 76,000 manuscripts in multiple languages including but not limited to Latin, Hebrew, Greek and Arabic [15].

[6, 16–18] are more common online datasets along with their web addresses. These online resources contain images from manuscripts written mainly in Latin. The biggest drawback of these resources is that, images need to be downloaded one at a time and even then, not all images are useful for testing algorithms. Therefore, these images need to be sorted and selected once they are downloaded. This is a time-consuming and tedious process. There are a few datasets available that can be used for manuscript author identification; [7, 10, 19, 20] are some of the major ones. Table 2 shows these datasets and their information.

The main issue with these datasets is that with the exception of the MPS dataset [7], other datasets do not provide proper dating information for their manuscripts. In addition, the Syriac [10] and IBN SINA [19] datasets focus on characters and are more suited for word detection, word segmentation and word annotation. MPS dataset is the most similar to KERTAS dataset with the major difference being the language of the manuscripts. In fact, MPS was the template on which KERTAS dataset was designed. The aim was to provide a dataset similar to MPS dataset for Arabic language that is able to support design, development, training

Table 2 Datasets comparable to KERTAS

Name	Language	Size
Syriac character [10]	Syriac	60 K characters
IBN SINA [19]	Arabic	51 folios, 20722 CCs
Barcelona Historical Marriages dataset (BH2 M) [20]	Spanish	244 book, 174 images
Medieval Paleographic Scale (MPS) [7]	Medieval Dutch	2858 charters
KERTAS dataset	Arabic	2505 images, 135 books

and testing of automatic manuscript age detection algorithms for Arabic historical documents.

4.2 KERTAS dataset

KERTAS dataset was designed particularly to assist in the effective training and testing of algorithms for document writer and age detection. That being said, the dataset is diverse and large enough to be equally useful in testing other algorithms such as document segmentation algorithms, text line and word extraction algorithms. The images selected to be the part of the dataset are also selected from a diverse set of documents ranging from manuscripts on mathematics, physics, Islamic history, metaphysics, etc. This diverse nature of these manuscripts provides some unique and challenging
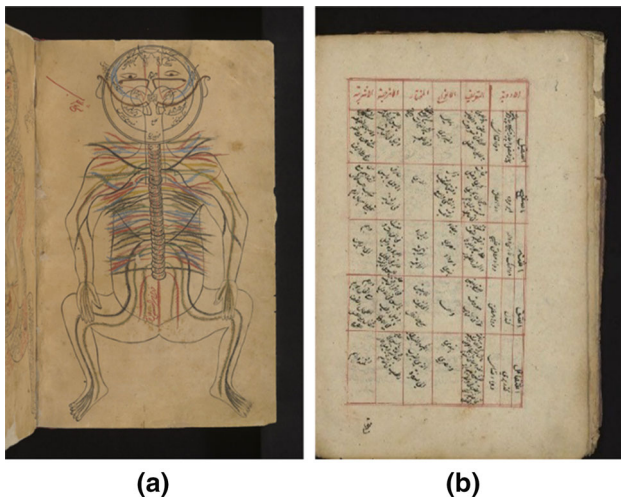
Fig. 6 a Example page with figures and **b** example page with tables, from Wellcome Library

images which can be used to test the limits of the algorithms being tested. Some images from manuscripts on mathematics contain drawings of figures and shapes (Fig. 6a shows an example), while others contain tables and lists within the text (an example is shown in Fig. 6b). These figures and tables are generally drawn with different color ink that are lighter than the text. In addition, the dataset also contains images of pages with comments on the margins written by different authors and different writing styles. Figure 6a, b shows examples of both cases. These images can be particularly challenging for writer identification algorithms and can help in designing robust writer identification algorithms.

5 Features extraction

Deciding on the most suitable feature extraction method is perhaps the most crucial step to achieve a high recognition rate. Age detection of historical manuscripts is a very challenging problem. We have to contend with the complexities inherent in working with noisy images, and there is an additional challenge we have to tackle. The class boundaries between the documents written in two adjacent centuries are highly nonconvex and nonlinear. This means that for documents written in two centuries and for documents written by two different authors, a very high interclass similarity will exist. In addition, handwritten documents by different authors are present in a single century, thus providing very high interclass variability.

5.1 Sparse representation-based approach

It is interesting to note that a similar kind of interclass similarity and intraclass variability is present in the domain of facial recognition.

Wright et al. [21] employed sparse representation for facial recognition. In this paper, we are following the same method with a small adjustment to consider all possible similarity of the test image in a single class (century) and across all classes while selecting the minimum number of training images required to represent each test sample adaptively.

This sparse representation provides new insight into the role of feature extraction and occlusion. This theory of compressed sensing (a technique of finding a sparse solution to an underdetermined linear system) suggests that the correct choice of feature space is no longer critical, however, giving a chance to a random feature to suitably represents a test image.

The sparse representation-based classification algorithm is in essence a nearest subspace selection algorithm. It uses l1 or l2 normalization as an optimization approach to select the nearest subspace to the document/image being evaluated. In simplest terms, the algorithm works as follows.

Given n_i training samples of the ith century, the matrix A_i can be presented as

$$A_i = [v_{i,1}, v_{i,2}, \ldots, v_{i,n}] \tag{1}$$

any new image from the same century will roughly be placed in the linear span of the training samples associated with object i:

$$y = \alpha_{i,1} v_{i,1} + \alpha_{i,2} v_{i,2} + \cdots + \alpha_{i,n} v_{i,n} \tag{2}$$

A new matrix A is defined as a concatenation of the n training images of all k centuries.

$$A = [A_1, A_2, \ldots, A_k] \tag{3}$$

and the linear representation of y can be defined as

$$y = A x_o \tag{4}$$

where

$$x_0 = [0, \ldots, 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,n}, 0, \ldots, 0]^T \tag{5}$$

is a coefficient vector with zero entries except those associated with the ith century. Equation (4) represents an underdetermined sparse linear system; the aim of this approach is to solve this system equation for x_0.

5.2 Handwriting style-based features

The similarity of handwriting style between manuscripts from the same period suggests considering classifying these document as a writing style-based classification problem. To evaluate sparse representation-based approach, we compared

it with some of the state-of-the-art writing style-based features that have been used in multiple researches [22–25]. These features are run-length feature as it was examined in [22], edge hinge and edge direction distribution as they were studied by Bulacu and Schomake in [23].

Run length is a multi-scale run feature that is obtained from the probability distribution of black and white pixels of a binary image [22]. Run-length feature is calculated after scanning the image into four directions: horizontal, vertical, left-diagonal and right-diagonal. Subsequently, probability distribution is estimated from normalized histogram of the scanned values. The approach is thoroughly explained in [22].

Edge hinge is obtained by calculating normalized histogram of curvature edge of the text. While edge direction is calculated from normalized histogram of text direction [23]. Both Edge hinge and Edge direction have been used in writer style identification written in different languages such as [22, 24, 25].

6 Results, analysis and discussion

The proposed approach presented above is evaluated on KERTAS dataset in order to show its performance. The experiment is set up by keep around 32% of the images selected randomly from the dataset as unseen test data. Every attempt has been made to keep this selection process random with only check been to make sure that all the classes are properly represented. The remaining images are kept as part the evaluation dataset.

Different image sizes are created by scaling the image to smaller sizes and no cropping takes place. We used the whole image because we are interested in using both the writing style and the layout style for document age detection as both elements are able to provide some information about the age of the document and it makes sense to utilize them both in order to improve the performance.

Different image sizes are tested to select the optimal feature set size. The rationale behind reducing image size is two folds: Firstly, reducing the image size reduces the dimensionality of the data thus producing a better underdetermined linear system which in term provided a better optimization solution. Secondly, the smaller variations such as noise and blemished are automatically removed thus providing an inherent robustness to the whole process. There is an issue with reducing the image size, i.e., if a manuscript image becomes too small, a lot of handwriting features that are useful in detecting the age of the manuscript may be lost. This reduces the possibility of a correct match. On the flip side, increasing the image size increases the dimensionality and

Table 3 Evaluation results for the sparse representation-based manuscript age detection algorithm

Image size	Accuracy with predefined folds (%)	Accuracy with random Train/Test split
12×12	80.62	28.46
25×25	92.00	41.11
50×50	**94.77**	**42.31**
100×100	92.62	42.31
200×200	92.32	41.56
250×250	92.32	41.41

Table 4 Writing style-based features results on KERTAS Dataset

Feature extraction method	Accuracy with predefined folds	Accuracy with random train/test split
Run length [22]	88.57%	85.71%
Edge direction [23]	70.48	66.66%
Edge hinge [23]	73.33	71.40

thus reducing the chance of a better optimization solution. In order to identify the optimal size for manuscript age detection, we tested the algorithm multiple images sizes starting from 12×12 (which was deemed optimal by authors in [26] for facial recognition) to 25×25, 50×50, 100×100, 200×200 and d 250×250. We also tested the algorithm with two different splits for the dataset. The first split is with predefined training and testing samples (95% training and 5% testing). This split enables us to train the algorithm with maximum variance of data. The second split of the dataset uses another approach where two-third random samples used for training (68%) and one-third for testing (32%). The results of the experiments are provided in Table 3.

The results show that as we increase the image size, initially the accuracy tends to improve and this is because more discriminative features are made available with the increase in size. The highest accuracy is obtained at the image size of 50×50. However, if we continue to increase the image size the accuracy tends to drop.

To evaluate the previous approach, we compare the results from sparse representation-based with some of state-of-the-art methods for writing style detection. First, the dataset was preprocessed to be used with writing style-based features. Text area was segmented and binarized using Otsu threshold method [27]. Afterward, features were extracted by run length, edge hinge and edge direction methods. We evaluated the performance of the features using k-Nearest Neighbor (k-NN) with $k = 3$. Table 4 presents the results of using writing style-based features on KERTAS dataset.

7 Conclusions

In this paper, we presented a dataset (KERTAS Dataset) accumulated specifically to assist researchers working on designing solutions and algorithms for digital paleography. The dataset consists of over 2000 high-quality, high-resolution digital images acquired from multiple historical handwritten Arabic manuscripts from multiple sources. Detailed metadata are provided for each image to assist in testing and verification of manuscript author detection and manuscript age detection algorithms. In addition, we presented a sparse representation-based approach to detect the age of manuscripts in order to highlight the suitability of the dataset. The algorithm also provides a baseline accuracy measure that can be compared with other algorithms developed in the future by using KERTAS dataset. Furthermore, we employed some writing style-based features to compare with the proposed approach and to study the consistency of writing style in each century in KETRAS dataset. The dataset will be made publicly available for research purpose through competition and IAPR TC-11 site.

Compliance with ethical standards

References

1. Falagas, M.E., Zarkadoulia, E.A., Samonis, G.: Arab science in the golden age (750–1258 C.E.) and today. FASEB J. **20**(10), 1581–1586 (2006). https://doi.org/10.1096/fj.06-0803ufm
2. Bellier, C.: Science-based Dating in Archaeology, L'Antiquité Classique 62:512, Brussil (1993). http://www.jstor.org/stable/41657909
3. Metzger, B.M.: Manuscripts of the Greek Bible: An Introduction to Palaeography. Oxford University Press, Oxford (1981)
4. Yeandle, L.: The evolution of handwriting in the English-speaking colonies of America. Am Arch **43**(3), 294–311 (1980)
5. Rehbein, M., Sahle, P., Schassan, T.: Codicology and Palaeography in the Digital Age, vol. 1. BoD–Books on Demand, Norderstedt (2009)
6. Aiolli, F., Ciula, A.: A case study on the system for paleographic inspections (SPI): challenges and new developments. Comput. Intell. Bioeng. Essays Mem. Antonina Starita **196**, 53–66 (2009). https://doi.org/10.3233/978-1-60750-010-0-53
7. He, S., Sammara, P., Burgers, J., Schomaker, L.: Towards style-based dating of historical documents. In: Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pp. 265–270. IEEE (2014)
8. He, S., Samara, P., Burgers, J., Schomaker, L.: A multiple-label guided clustering algorithm for historical document dating and localization. IEEE Trans. Image Process. **25**(11), 5252–5265 (2016). https://doi.org/10.1109/TIP.2016.2602078
9. Wahlberg, F., Mårtensson, L., Brun, A.: Large scale style based dating of medieval manuscripts. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, pp. 107–114. ACM (2015)
10. Howe, N.R., Yang, A., Penn, M.: A character style library for Syriac manuscripts. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, pp. 123–128. ACM (2015)
11. Abulhab, S.D.: Roots of modern Arabic script: From Musnad to Jazm, Sawt Dahesh. 50–51 (2007–2009)
12. Foundation A-FIH: Islamic codicology: an introduction to the study of manuscripts in Arabic script. Al-Furqān Islamic Heritage Foundation (2006)
13. At, E.: Arabic Manuscript a Study in the Dimensions of Time and Place. Syrian Cultural Ministry, Damascus (2011). **(in Arabic)**
14. Gacek, A.: Arabic Manuscripts: A Vademecum for Readers, vol. 98. Brill, Leiden (2009)
15. Le Bourgeois, F., Kaileh, H.: Automatic metadata retrieval from ancient manuscripts. In: Document Analysis Systems VI, pp. 75–89. Springer, Berlin (2004)
16. Feuerverger, A., Hall, P., Tilahun, G., Gervers, M.: Using statistical smoothing to date medieval manuscripts. In: Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen, pp. 321–331. Institute of Mathematical Statistics (2008)
17. Fecker, D., Asi, A., Pantke, W., Märgner, V., El-Sana, J., Fingscheidt, T.: Document writer analysis with rejection for historical arabic manuscripts. In: Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, pp. 743–748. IEEE (2014)
18. Garain, U., Parui, S., Paquet, T., Heutte, L.: Machine dating of handwritten manuscripts. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 759–763. IEEE (2007)
19. Farrahi Moghaddam, R., Cheriet, M., Adankon, M.M., Filonenko, K., Wisnovsky, R.: IBN SINA: a database for research on processing and understanding of Arabic manuscripts images. In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 11–18. ACM (2010)
20. Fernandez-Mota, D., Almazan, J., Cirera, N., Fornes, A., Llados, J.: BH2M: the Barcelona Historical Handwritten Marriages database. In: 22nd International Conference on Pattern Recognition (ICPR), Swedish Soc Automated Image Anal, Stockholm, Sweden, August 24–28 2014. International Conference on Pattern Recognition, pp. 256–261. IEEE Computer Society, Los Alamitos. https://doi.org/10.1109/icpr.2014.53 (2014)
21. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 210–227 (2009)

22. Djeddi, C., Siddiqi, I., Souici-Meslati, L., Ennaji, A.: Text-independent writer recognition using multi-script handwritten texts. Pattern Recognit. Lett. **34**(10), 1196–1202 (2013)

23. Bulacu, M., Schomaker, L.: Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 701–717 (2007). https://doi.org/10.1109/TPAMI.2007.1009

24. Brink, A.A., Smit, J., Bulacu, M.L., Schomaker, L.R.B.: Writer identification using directional ink-trace width measurements. Pattern Recognit. **45**(1), 162–171 (2012). https://doi.org/10.1016/j.patcog.2011.07.005

25. Siddiqi, I., Vincent, N.: Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. Pattern Recognit. **43**(11), 3853–3865 (2010). https://doi.org/10.1016/j.patcog.2010.05.019

26. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. **3**(1), 71–86 (1991)

27. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979). https://doi.org/10.1109/TSMC.1979.4310076