ISSN 2278-3091
Volume 8, No.1.3, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse4881.32019.pdf
https://doi.org/10.30534/ijatcse/2019/4881.32019

State of the Art in Digital Paleography

Nur Atikah Arbain[1]*, Mohd Sanusi Azmi[1], Azah Kamilah Muda[1], Maslita Abdul Aziz[1], Intan Ermahani A. Jalil[1]
[1]Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian
Tunggal, Melaka, Malaysia, nuratikah9.arbain@gmail.com

ABSTRACT

Digital paleography is an approach used to assist paleographers in deciding the origin of manuscripts. This is done by recording types of writings present in old manuscripts. It uses digital representation of book hands as a tool to support paleographical analyses by human experts. There are six types of manuscripts selected which are Arabic, Chinese, Jawi, Indian, Latin and Roman. These types of manuscripts are discussed through their current contribution in the digital paleography field. The main purpose of this paper is to discuss the current work on digital paleography for selected types of manuscripts. Thus, we identified the approaches and methods used to define the types of handwritings in old manuscripts

Key words : Ancient Manuscript; Digital Paleography; Feature Extraction; Paleography, Handwriting.

1. INTRODUCTION

Paleography is the study of identifying manuscript origin details such as originality, writing date, provenance, as well as a number of authors. Most of the ancient manuscripts such as Arabic, Chinese, Hebrew and Latin were old and corrupted due to poor of storage conditions and quality of written parchment. However, with advanced technological now, it aids paleographers in studying ancient documents.

Each of languages has their own difficulties in identifying the pattern, shape, styles and type of handwriting. For example, Arabic document has been stated as one of challenging languages because of Arabic characters were hard to recognize. Most of the characters have dots and some of Arabic text were complex to segment due to stroke and shape of Arabic characters [1]. The similar problem also occurs for Chinese document. The complex shape and strokes become a challenging task to researchers in identifying the handwriting in ancient manuscripts [2].

In recent years, the field of digital paleography has grown with a dynamic response due to paleographers who seem to be subjective with their views. With technological advances, research has been made to be more reliable. Digital paleography helps experts in deciding the manuscript origin by scoring the type of writing of an old manuscript [3]. The purpose of digital paleography described by [4] is to use the digital representation of book hands as a tool to support paleographical analysis by human experts. It describes certain graphics style of handwriting and compares it with different scripts in relation to its geographical and chronological aspects. Most research works aim to recognize the type and style of writing as well as paleography. Among the research works, identifying the writers have greater values compared to the other details.

We present this paper with the intention of discussing the state of the art in digital paleography. This paper is organized as follows: Section 2 presents a brief of paleography background. Section 3 discusses state of the art in digital paleography and finally, Section 4 summarizes this paper.

2. BACKGROUND OF PALEOGRAPHY

The digital form of paleography only began in 1999 where the studies only focused on Roman, Hebrew and Hindi domain. The System for Palaeography Inspector (SPI) was first developed in 1999 by researchers from University of Pisa. Nevertheless, this system was incomplete due to the insufficient features and the system does not support Windows XP operating system [4]. In 2009, a case study about SPI was reported by same authors [5]. The case study was about to update their research work from various aspects.

In 2004, a preliminary study in Hebrew manuscript was conducted by [6]. However, the study only used five Hebrew alphabets where no clear justification was stated over the choice of five alphabets. The proposed technique by [6] only revolved around the range of spaces within the alphabets. Then, in 2005, [7] has proposed an input sensitive thresholding algorithm for ancient Hebrew manuscripts.

Two years later, the studies in the Roman manuscript were explored. [8] has applied a global approach which was a statistical method. The statistical method which was based on Haralick characteristics were used to obtain 12 features for identifying the types of Roman script. Nevertheless, the method used was only able to determine the dominant script in Roman manuscripts.

With the limitations of methods used by previous researchers as aforementioned, the studies in digital paleography have demanded a different method that can be used to identify the type of handwriting and origin of the script.

3. STATE OF THE ART IN DIGITAL PALEOGRAPHY

As of today, there is numerous project works being scrupulous with evolving such methodologies. These methodologies have been implemented in preparing to digitize the paleography into digital form. [9] introduces the computerized paleography through two tools which were developed for aiding the paleographers in paleography. The handwriting matching tool and paleographic classification tool were designed to provide a high level of accuracy and concise justification of the inferred results [9]. However, these methods were stated has only tested based solely on Hebrew (Genizah) handwriting.

3.1. Digital Jawi Paleography

In meantime, with a vast number of Jawi manuscripts in Malaysia, a preliminary work on digital Jawi paleography has been introduced in 2010 with reports of discoveries and paper publications [3], [10], [11]. A framework of digital Jawi paleography was proposed in the earliest phase of the development. It was used to identify the types of Arabic calligraphy in Jawi manuscripts [3]. The research in digital Jawi paleography has been expanded by [12] uses a triangle geometry method to extract features from Arabic calligraphy. The analysis on digital Jawi paleography was performed through Arabic calligraphy dataset. The study has reported that the proposed technique using triangle geometry method was successfully implemented in digital paleography of Jawi characters.

Nevertheless, no Jawi character database was available to be reached even many types of research have been done. Thus, in 2015, the first printed Jawi database was proposed by [13]. The printed Jawi database namely Database Printed Jawi (DPJ) was developed to assist the researchers in analysing and evaluating Jawi character recognition method. The Jawi database contains four types of fonts which have 1524 characters and also 168 printed words and sentence images [13]. The DPJ dataset has been used by [14] in their study. [14] has built optical character recognition using offline and online application which used to compare the performance between them based on their execution time. The results have shown online application executes faster compare to the offline application.

3.2. Latin Paleography

The study in digital paleography also has been evolved in Latin manuscript. Begin with first digital paleography in 1999 [4], many types of research have been performed until now. [15] uses convolutional neural networks in identifying medieval Latin manuscripts in their studies. The study of [15] contributes to the Latin paleography in recognizing the type of Latin scripts. By using the CLaMM (Classification of Latin Medieval Manuscripts), the studies offered the visualizations which were used to cater the issues of the black box in Digital Paleography [16], [17]. Besides that, [18] also has used CLaMM dataset in their studies. They had performed an analysis on CLaMM dataset and also on 40 Arabic computer fonts using the convolutional neural network. The proposed method focused on data augmentation which had produced a good result by achieving the highest accuracy of 86.6% in predicting the paleographic scribal script on Latin manuscripts.

3.3. Arabic Paleography

Paleography in Arabic manuscripts has been discussed from date to date. Most Arabic paleographical studies have been largely based on existing dated samples with the dates aiding as landmarks. The contributors and editors alike show a keen appreciation of the significance in the historical development of Arabic paleography and calligraphy. Though others have approached this issue differently, some having more familiarity than others with the historical, less artistic, but more difficult and fundamental phase in the study of Islamic writing [19]. The Arabic paleography can be associated with Malay paleography because the Jawi scripts were created based on Arabic characters while the orthography system of Malay paleography was strongly influenced by the Arabic language [20]. However, no serious research for Arabic paleography has been conducted until now. Most research for Arabic scripts was found only discussing its origins and calligraphy [20].

The task to recognize the Arabic handwriting becomes more challenging amongst researchers alike when dealing with dots in, on and in between of characters. According to [1], Arabic character recognition can be considered as a challenging task compared to other handwritings such as Latin, Japanese and Chinese. Besides, there are very limited works reported for Arabic handwriting recognition [1]. A new recognition system has been proposed by [21] to inspect the feature extraction method for the recognition of handwritten Arabic character. The main goal is to exploit the recognition rate with the minimum volume of elements [21]. Similar studies in [1], [21] also collect the Arabic characters. However, the way of study in [1] to obtain data input was from various writers while the studies [21] had obtained data input from people in different age groups. Even so, both studies in [1], [21] processed the data input manually by scanning the form they had collected from one by one. The [21] had used a form that was written in different colours such as blue and black by the writers.

3.4. Chinese Paleography

However, in Chinese paleography, it more concerned with empires far mediocre from the modern day. Paleographers work on inscriptions and other materials endlessly, struggling to bring the information and bonds of those past times into the present day. A digital system for paleography must be able to handle inconsistency in both dimensions, flawlessly and accurately. If this is not done, information will be lost in the exchange and research findings will fail as a result. According to [22], a digital system for paleography must provide the outline for creating significant relations among different

written editions and provide a method to relate various understandings of the basic data.

There are seven stages featured in Chinese advanced paleography which can be alluded in [22]. The investigation of Chinese paleography has been examined by [2]. Various uncovered compositions including uncommon versions of significant Chinese works of art and missing writings portrayed different parts of Chinese culture and conventions [2]. Chinese paleography takes a shot at how the antiquated Chinese engravings are elaborately, fundamentally and tastefully introduced by the Chinese calligraphy. [2] had expressed that the interdisciplinary investigation of Chinese paleography, calligraphy, and example acknowledgment had given a progressed computerized way to deal with the comprehension of uncovered antiquated Chinese reports and upheaval of customary Chinese paleography.

The investigation of Chinese paleography, calligraphy, and example acknowledgment which conjointly examines the expressive variety and auxiliary multifaceted nature of old Chinese contents develop an imaginative digital-based comprehensive Chinese paleography. The exploration in [2] explains in detail of Chinese paleography, calligraphy and example acknowledgment which spotlights on the styles and contents in unearthed antiquated Chinese reports. Other than that, [2] likewise referenced that the Transparent Transcription approach was a viable and creative way to deal with learn Chinese paleography. Also, this comprehensive methodology speaks to the future that can't be maintained a strategic distance from in investigating the Chinese paleography in the computerized period [2]. The exploration in [2] has detailed that the Transparent Transcription is compelling and best in class way to deal with concentrate Chinese paleography. The investigation of Chinese penmanship acknowledgment has gotten requesting graciousness for a considerable length of time. Subsequently, the investigation of Chinese penmanship has been investigated among specialists in the penmanship acknowledgment field.

3.5. Roman Paleography

The study of Roman handwriting has been in existence for more than four decades [23]. There were no available standard datasets that could be reached by the research community at that time. However, the research and development of digit recognition have gone on a rapid advancement since the last decade. Thus, the study of Roman handwritings has been getting attention and explored among researchers of handwriting recognition. The Roman dataset, MNIST, was developed in 1992 and was known as the National Institute of Standard Technology (NIST) at the beginning. Then, the dataset was named MNIST [24]. The MNIST dataset has been widely used among researchers in handwriting recognition [23], [25].

The research in Roman handwriting [25], [26], the researchers have used different methods to obtain the data input for their research. Moreover, there are also two types of datasets that were used in [25]. These datasets are known as contemporary handwritings which are Firemaker [27] and IAM [28]. The Firemaker [27] consists of 1004 pages where four pages were written by every 251 students. From the 4 pages, only pages 1 and 4 were used in the research. Whilst, in the IAM dataset, it consists of English handwriting that has been written by 657 subjects using different pens [28]. In 2010, the researchers [26] had also used the IAM data as the data input in the research. 956 words of the offline IAM dataset have been trained whereas each word is represented by 29 statistical features.
[26] has presented the swarm intelligence as a features weighting mechanism to distinguish between a feature that has high importance and low importance in the identification process. Then, the weights are acquired from swarm experiments which are then used to alter the feature scores. It will then be used to recognize the most vital subset feature for the writers' classification process. The Binary Particle Swarm Optimization (BPSO) is used as a feature selection method whereas the Euclidean Distance (ED) is used as an evaluation function for BPSO. However, [25] has used Quill and Quill-Hinge which are known as pen-dependent features in their experiments.

3.6. Indian Paleography

Indian is known as a multilingual country of more than one billion people residing in the country with 22 constitutional languages and 10 different scripts. Bangla happens to be the most popular language and a script of the Indian sub-continent which is Devanagari [29]. This language has been used by more than 200 million people around the world. In addition, Bangla is also known as the official language of Bangladesh [29]. The Bangla script is similar as two other Indic scripts which are Viz., Assamese and Manipuri [29].

The study of Indian paleography grew out of essential for interpreting a large number of writings found in the sub-continent. With advanced technological now, research and development of handwritten character recognition methodologies for Indian scripts have improved immensely. Based on [30], the nonexistence of standard or benchmark databases has become a major constraint to the research on handwritten character recognition of Indian scripts. However, several standard databases such as NIST, MNIST, CEDAR and CENPARMI are available for Latin numerals [30]. In [30], the researcher had presented a pioneering effort towards the development of handwritten numeral databases of Indian scripts.

Nevertheless, a complete handwritten numeral database of Bangla has been developed in 2006 at the Computer Vision and Pattern Recognition Unit Laboratory of Indian Statistical Institute, Kolkata [29]. This database is known as the ISI database [29]. The current database has several components which include both online and off-line handwritten numerals. The samples of numeral strings and isolated numerals have been gathered under both modes of writing. The printed Indian scripts; Devanagari [31] texts and Bangla OCR System [32] are the significant works being done on recently [29]. However, researchers for handwritten characters on several Indian scripts such as Bangla [32], [33] and Devanagari [31]

are lacking. According to [29], the source of different databases collected is either in laboratory surroundings or from minor clusters of the concerned population. The database was developed to make fellow researchers feel free for the persistence of handwriting recognition research on Indic scripts.

The research to identify the Telugu, Devanagari and English manuscript by using discriminating features was introduced by [34]. The aim of [34] was to identify the differing manuscript regions of the document in order to support the document to the Optical Character Recognition (OCR) of an individual language. The OCR is a technology that enables researchers to convert different types of documents such as scanned paper documents into editable and searchable data [35]. The researchers in [34] have developed a model to recognize and detach the text lines of Telugu, Devanagari and English manuscripts from the printed trilingual document.

In Indian paleography [34], the researchers have proposed a new model to identify and separate the text lines of Telugu, Devanagari and English scripts. Therefore, the [34] has used the new proposed method to extract the distinct values from the top and bottom profile based features. Unlike the research in [34], the researcher [36] had used Optical Character Recognition (OCR) in Devanagari scripts only. This is because not much research has been done by using the OCR in Indian scripts such as Devanagari. Therefore, the researcher has come out with the proposed model for research in [36].

4.CONCLUSION

A brief review of state of the art in digital paleography by several languages has been discussed in this paper. A lot of effort has been spent in the study of digital paleography. Nevertheless, the study in digital paleography has provided a lot of benefits in identifying the ancient manuscripts. Besides, most ancient manuscripts are valuable. Even though the study of paleography has been explored long time ago, the lack of technologies becomes a major factor among researchers to study. Thus, this paper has been presented in general and specifically concentrating on related work of digital paleography for each of languages. The current trend on techniques that have been used by numerous researchers was aided in digital paleography.

ACKNOWLEDGEMENT

This work is financially supported by the FRGS grant: FRGS/1/2017/ICT02/FTMK-CACT/F00345 which under the Ministry of Higher Education Malaysia. Next, thanks to the Universiti Teknikal Malaysia Melaka and Faculty of Information Technology and Communication for providing excellent research facilities.

REFERENCES

1. J. H. Alkhateeb, **"Off-Line Arabic Handwritten Isolated Character Recognition,"** *Int. J. Eng. Sci. Technol.*, vol. 7, no. November, pp. 251–257, 2015.
2. X. Wen, **"Chinese Paleography, Calligraphy, and Pattern Recognition: Styles and Scripts in Excavated Ancient Chinese Documents,"** in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 951–956. https://doi.org/10.1109/ICDAR.2011.193
3. M. S. Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, and A. Abdullah, "**Digital paleography: Using the digital representation of Jawi manuscripts to support paleographic analysis,**" *Proc. 2011 Int. Conf. Pattern Anal. Intell. Robot. ICPAIR 2011*, vol. 1, no. June, pp. 71–77, 2011. https://doi.org/10.1109/ICPAIR.2011.5976914
4. A. Ciula, **"Digital palaeography : using the digital representation of medieval script to support palaeographic analysis,"** vol. 1, no. Spring, pp. 1–31, 2005. https://doi.org/10.16995/dm.4
5. F. Aiolli and A. Ciula, **"A case study on the system for paleographic inspections (SPI): Challenges and new developments,"** *Front. Artif. Intell. Appl.*, vol. 196, no. 1, pp. 53–66, 2009.
6. I. B. Yosef, K. Kedem, I. Dinstein, M. Belt-Arie, and E. Engel, "**Classification of Hebrew Calligraphic Handwriting Styles: Preliminary Results,**" *Proc. - First Int. Work. Doc. Image Anal. Libr. - DIAL 2004*, no. ii, pp. 299–305, 2004.
7. I. Bar-Yosef, "**Input sensitive thresholding for ancient Hebrew manuscript**," *Pattern Recognit. Lett.*, vol. 26, no. 8, pp. 1168–1173, Jun. 2005. https://doi.org/10.1016/j.patrec.2004.07.014
8. I. Moalla, F. Lebourgeois, H. Emptoz, and a. M. Alimi, "**Image analysis for palaeography inspection,**" *Proc. - Second Int. Conf. Doc. Image Anal. Libr. DIAL 2006*, vol. 2006, pp. 303–310, 2006.
9. L. Wolf, L. Potikha, N. Dershowitz, R. Shweka, and Y. Choueka, "**Computerized paleography: Tools for historical manuscripts,**" in *Proceedings - International Conference on Image Processing, ICIP*, 2011. https://doi.org/10.1109/ICIP.2011.6116481
10. Khairuddin Omar, Mohd Sanusi Azmi, S. N. H. Syeikh Abdullah, M. F. Nasrudin, and A. N. Azizi, "**Kerangka Paleografi Jawi Digital : Satu Cadangan Awal,**" *Semin. Teknol. Mklm. 2010*, pp. 1–14, 2010.
11. K. Omar, M. S. Azmi, M. F. Nasrudin, A. K. Muda, and C. W. S. B. C. W. Ahmad, "P**aelografi Jawi Digital,**" pp. 1–18.
12. M. S. Azmi and K. Omar, "**Features Extraction of Arabic Calligraphy using extended Triangle Model for Digital Jawi Paleography Analysis,**" *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 5, pp. 696–703, 2013.
13. K. Saddami, K. Munadi, and F. Arnia, "**A database of printed Jawi character image**," *Proc. 2015 3rd Int. Conf. Image Inf. Process. ICIIP 2015*, pp. 56–59, 2016. https://doi.org/10.1109/ICIIP.2015.7414740
14. Nazaruddin and S. Muchallil, "**Performance Comparison Online and Offline for Printed Jawi Character Recognition,**" *Indian J. Sci. Technol.*, vol.

10, no. 12, pp. 1–5, 2017.
15. M. Kestemont and D. Stutzmann, "**Script identification in medieval Latin manuscripts using convolutional neural networks,**" in *Digital humanities 2017. Book of abstracts (Montreal, August 10, 2017)*, 2017, pp. 283–285.
16. D. Stutzmann, "**Digital Palaeography: New Machines and Old Texts (Dagstuhl, 20-24 July 2014),**" vol. 4, no. 7, pp. 112–134, 2014.
17. T. Hassner, M. Rehbein, P. A. Stokes, and L. Wolf, "**Computation and Palaeography : Potentials and Limits,**" *Manif. from Dagstuhl Perspect. Work.*, vol. 2, no. 1, pp. 14–35, 2013.
18. C. Tensmeyer, D. Saunders, and T. Martinez, "**Convolutional Neural Networks for Font Classification,**" 2017. https://doi.org/10.1109/ICDAR.2017.164
19. N. Abbott, "**Review: Arabic Paleography**," vol. 8, pp. 65–104, 2011.
20. Wan Ali Wan Mat, "**Paleografi Jawi : Satu Pengenalan**," *J. Filol. Melayu*, vol. 14, pp. 65–69, 2006.
21. F. M. H. Al-shareefi, "**A Haar Wavelet-Based Zoning For Offline Arabic Handwritten Character Recognition**," no. 2, pp. 575–585, 2015.
22. R. S. Cook, "**Unicode Chinese paleography : making the evolutionary leap from bone , bronze , silk , and paper , to electronic bits,**" in *Early Chinese Workshop Series*, 2005.
23. M. S. Azmi, K. Omar, M. F. Nasrudin, B. Idrus, and K. Wan Mohd Ghazali, "**Digit Recognition for Arabic/Jawi and Roman using Features from Triangle Geometry,**" *AIP Conf. Proc.*, vol. 1522, pp. 526–537, 2013.
24. A. Borji, M. Hamidi, and F. Mahmoudi, "**Robust handwritten character recognition with features inspired by visual ventral stream**," *Neural Process. Lett.*, vol. 28, no. 2, pp. 97–111, 2008. https://doi.org/10.1007/s11063-008-9084-y
25. a. a. Brink, J. Smit, M. L. Bulacu, and L. R. B. Schomaker, "**Writer identification using directional ink-trace width measurements**," *Pattern Recognit.*, vol. 45, no. 1, pp. 162–171, 2012. https://doi.org/10.1016/j.patcog.2011.07.005
26. K. M. Bin Abdl and S. Z. M. Hashim, "**Swarm-based feature selection for handwriting identification**," *J. Comput. Sci.*, vol. 6, no. 1, pp. 80–86, 2010. https://doi.org/10.3844/jcssp.2010.80.86
27. L. Schomaker, Louis Vuurpijl, and Lambertus Schomaker, "**Forensic writer identification: A benchmark data set and a comparison of two systems,**" in *NICI*, 2000.
28. U.-V. Marti and H. Bunke, "**A full English sentence database for off-line handwriting recognition**," *Proc. Fifth Int. Conf. Doc. Anal. Recognition. ICDAR '99 (Cat. No.PR00318)*, no. November, pp. 705–708, 1999. https://doi.org/10.1109/ICDAR.1999.791885
29. B. B. Chaudhuri, "**A Complete Handwritten Numeral Database of Bangla – A Major India Script,**" 2006.
30. U. Bhattacharya and B. B. Chaudhuri, "**Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals,**" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 444–457, 2009. https://doi.org/10.1109/TPAMI.2008.88
31. V. Bansal and R. M. K. Sinha, "**Integrating knowledge sources in Devanagari text recognition system,**" *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 30, no. 4, pp. 500–505, 2000. https://doi.org/10.1109/3468.852443
32. B. . Chaudhuri and U. Pal, "**A complete printed Bangla OCR system**," *Pattern Recognit.*, vol. 31, no. 5, pp. 531–549, 1998. https://doi.org/10.1016/S0031-3203(97)00078-2
33. A. Dutta and S. Chaudhury, "**Bengali alpha-numeric character recognition using curvature features,**" *Pattern Recognit.*, vol. 26, no. 12, pp. 1757–1770, 1993. https://doi.org/10.1016/0031-3203(93)90174-U
34. M. C. Padma and P. a. Vijaya, "**Identification of Telugu, Devanagari amd English Scripts using Discriminating Features,**" *Int. J. Comput. Sci. Inf. Technol.*, vol. 1, no. 2, pp. 64–78, 2009. https://doi.org/10.1080/18756891.2008.9727609
35. "**What is OCR and OCR Technology**." [Online]. Available: http://www.abbyy.com/finereader/about-ocr/what-is-ocr/. [Accessed: 03-Nov-2015].
36. T. R. Zalke and P. V. N. Bhonge, "**An Optical Character Recognition System for Indian Scripts**," no. 5, pp. 392–394, 2015.