Journal of Ayurveda & Holistic Medicine

https://jahm.co.in

eISSN-2321-1563

REVIEW ARTICLE OPEN ACCESS

ARTIFICIAL INTELLIGENCE POWERED OCR MODELS FOR DIGITIZING AYURVEDA MANUSCRIPTS

HARITHA T J[1]*, SOURAV K T[2], RESMI B[3]

[1]*2nd year PG scholar, [3]Professor and HOD, Department of Samhitha Samskrita & siddhanta, Government Ayurveda College, Trivandrum.

[2] 2nd year PG scholar, Department of Shalyatantra, Government Ayurveda college, Trivandrum.

Corresponding Author Email: harithajayaraj96.hj@gmail.com Access this article online: https://jahm.co.in

Submitted on- 07-09-24	Revised on- 24-09-24, 03-10-24	Accepted on-05-10-24

ABSTRACT:

The enormous amount of unexplored ayurveda manuscripts has to undergo the steps of manuscriptology before it is to be made available as refined accessible documents or books. Transcription is one among the steps of manuscriptology, which is a very time consuming and tedious process. Researchers are required to write or transcribe, alphabet to alphabet, from a manuscript and convert it into a digitally editable format. The Optical Character Recognition (OCR) software, websites and applications powered by artificial intelligence and machine learning technology can provide a promising solution for this hurdle. The currently available software and applications are able to recognize and transcribe printed documents with astonishing accuracy, yet falls short in the accurate transcription of handwritten documents. Though some among them can be trained and customized, those software demands high expertise and technical skill. Thus, development of a user-friendly interface, where the scholars initially feed samples of a specific manuscript and further training it by providing samples for confusing or indistinct characters, will enable them to create an OCR model developed for the selected specific manuscript, which is capable of transcribing it with at most accuracy. This proposed method, by significantly reducing the time required for the transcription of manuscripts and by minimizing the required level of technical expertise, can bring out the knowledge form the unexplored Sanskrit ayurveda manuscripts and thereby accelerate researches in ayurveda.

Keywords: Optical Character Recognition, OCR, Ayurveda Manuscripts, Handwriting Analysis, Artificial Intelligence, Machine Learning, Manuscript Digitization.

INTRODUCTION

Ayurveda, the ancient Indian system of medicine, was documented in numerous manuscripts dating back to about 1st to 2nd century CE. India is considered as the biggest repository of manuscripts accounting to a collection of 10 million manuscripts, among them only 20,000 ayurveda manuscripts [1] are identified but many are unexplored [2] These manuscripts are invaluable repositories of knowledge, containing insights into herbal remedies, diagnostic techniques, and therapeutic practices.

Manuscriptology is the scientific or structural study of hand written documents credited with fair antiquity [3]. Manuscriptology involves primary and secondary steps. Collection, conservation and cataloguing constitute the primary steps while transcription, translation, critical edition and publication constitute the secondary steps.[4] Among these, transcription is a tedious and time-consuming process. Traditional methods of manuscript transcription involve painstakingly transcribing each character, a process that can take weeks or even months for a single manuscript. Scripts are the characters used to symbolize and denote the language. *Brahmi, kutila, kharoshti, gupta* and *grantha* are the various ancient scripts. Among them Brahmi script was widely used throughout India. Later the *brahmi* system diverged in to northern and southern systems. Devanagari and Gujarati were in northern *brahmi* while *vattezhuth*, Kannada, Malayalam etc. comes under the southern *brahmi.* [5] Though, huge amount of ayurveda manuscripts which are available in Sanskrit are written in the Devanagari script, many regionally available ayurveda Sanskrit manuscripts are written in regional scripts like Kannada, Malayalam etc. For example, most of the ayurveda manuscripts in Oriental Research Institute Mysore are in Kannada script while the verses are Sanskrit. Similarly, those preserved in the Cambridge University library and British library are in Nepalese and Sinhalese respectively.[6] These diverse scripts used in Ayurveda manuscripts pose additional challenges for accurate transcription.

Optical Character Recognition (OCR) technology [7] offers a promising solution to transcription of ayurveda manuscripts. By automatically extracting text from images, OCR can significantly reduce the time and effort required for transcription. OCR software and applications recognizes the printed text with astonishing accuracy [8]. However, conventional OCR software often struggles with handwritten documents [9], leading to errors and inaccuracies in transcription.

This article proposes a novel approach to OCR tailored specifically for Sanskrit Ayurveda manuscripts. By incorporating principles of forensic handwriting analysis and leveraging machine learning techniques, OCR models capable of accurately transcribing handwritten manuscripts can be developed.

Current OCR technologies

Optical Character Recognition (OCR) is an emerging technology which is extensively made use of digitization of documents in various fields. Current OCR software have been endowed with

the capability of accurately detecting the text as well as various characteristics of the provided documents including complex content like tables, equations, footnotes, headers, layout etc. Very few OCR software which are incorporated with machine learning models and neural networks are able to recognize handwritten documents or annotations made on printed files to some extent with arguable accuracy.

The most popular and advanced OCR software and applications can be categorized into open source and premium, each with different capabilities and features suitable for various uses.

Open-source OCR software are Tesseract OCR, OCRmyPDF, Calamari OCR, EasyOCR, Kraken OCR etc.

Premium OCR software are ABBYY FineReader, Adobe Acrobat Pro DC, Readiris, Nuance OmniPage, Google Cloud Vision API, Microsoft Azure Computer Vision etc.

Tesseract, the open-source tool powered by google which have options for customization make it ideal for research applications as well as general text recognition from documents. Yet ample amount of technical skill is required for the effective implementation. On the other hand, premium software like ABBYY FineReader, Adobe Acrobat Pro DC, Readiris, and Nuance OmniPage provides users with better features, higher accuracy and enhanced efficiency. Large scale OCR can be performed in cloud-based services such as Google cloud vision API and Microsoft Azure Computer Vision which helps in analysis of extensive data thereby increasing the productivity.

Sanskrit OCR software

Even though OCR technology for Sanskrit language have made significant progress, still it faces various challenges due to diversity of the Sanskrit language and the complex nature of the scripts like Devanagari. The quality of the document or scanned image even if it is a printed document with uniform font, evokes difficulty in differentiation between the alphabets like *pa(प), ya(य), ba(ब),* and *va(व).* Apart from the basic alphabets in the Devanagari script, it consists of large set of characters, including ligatures and conjunct consonants which further increases the difficulty in recognition. The ancient wisdom of Ayurveda preserved in centuries old manuscripts are written on various materials like palm leaf, cloth, leather or paper using different tools are often damaged which reduces the accuracy of OCR.

Tesseract OCR supports Devanagari and can be customized to attain better accuracy but struggles with the particulars of Sanskrit language and requires elaborate training and pre-processed data. Google cloud vision OCR has utility for large scale projects but identifies Sanskrit, only with moderate accuracy and has a higher subscription cost. The mobile friendly Sanskrit OCR apps provide quick scanning and even translations but cannot provide advanced features and customization options of desktop or cloud-based services.

There are several websites to OCR Sanskrit like the Indic OCR Project. It is constructed with a web-

based interface where the images of printed text can be uploaded and a corresponding output is received. But it processes only good quality images of printed documents and has a reduced accuracy in complex as well as poorly scanned documents. Another website is the Sanskrit OCR by Sanskritworld. Even though it is designed for recognizing Sanskrit documents in Devanagari the accuracy in OCR for handwritten documents are questionable and also it lacks customization options or training capabilities. Google Drive OCR which is a cloud- based service which supports Devanagari script. Even though it is an integrated OCR function within google drive and allows easy sharing and collaboration, it lacks accuracy for manuscripts due to the lack of specific tailoring for Sanskrit. Transkribus, on the other hand is highly used for literary research involving digitization of Sanskrit manuscripts, supports user fed data for training the OCR model but its steep learning curve and subscription cost makes it less feasible.[10] i2OCR is a free online OCR service which supports Devanagari. It is rather a quick and convenient tool for small documents but highly depends on image quality, and fails in recognizing hand written documents and complex manuscripts.

The limitations of these websites or apps are the lack of accuracy when it comes to the handwritten documents and the need of skill in technical aspects. The literary research field of ayurveda mostly deals with handwritten, damaged and old manuscripts which further reduces the accuracy in OCR. Development of a trainable model using deep learning techniques and neural networks with high accuracy in recognizing handwritten document and a user-friendly interface will help the researchers in greatly speeding up the process.

METHODOLOGY

The following are the steps of the proposed methodology.

Collection of manuscripts is the first and foremost step. It is a very crucial step as it provides the primary data for machine learning by offering diverse handwritten samples.

The various sources of manuscript repository in India and abroad include;

The Oriental research institutes around India are rich repositories of ayurveda manuscripts. For example: oriental research institute Mysore, oriental research institute and manuscript library in Kariyavattom, Bhandarkar Oriental Research Institute in Pune and Rajasthan Oriental Research Institute in Jodhpur etc. Libraries of various universities and colleges include: Lal Chand library, Saraswati bhavan library Varanasi, French library Pondicherry, Saraswati Mahal Library in Thanjavur etc. [11], Private collections and libraries of Asiatic society are also huge repositories of manuscripts. Britain is the largest repository of ayurveda manuscript outside India with manuscripts preserved at the Libraries of Oxford University, Cambridge University, British library, the Wellcome institute for history of Medicine etc.

Machine learning training: The collected samples of various Sanskrit ayurveda manuscripts with different scripts are fed to OCR models for training such as deep neural network to recognize the nuances and variations of handwriting styles. The

manuscripts are digitized and annotated with ground truth text which is the accurate and verified data used for the training of the OCR models. Through the models, images are enhanced, noises removed and segmented for the training and validation of test data. Feature extraction and sequence prediction can be carried out by Deep neural networks, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).[12] The model parameters are to be updated with forward passes, loss calculation and back propagation. The accuracy and efficiency of the developed model are to be assessed by metrics like word error rate (WER) and Character error rate (CER).[13]

Since there is consistency in the font of every alphabet of printed document OCR of the document is relatively quick and easy. But in the case of a manuscript there are natural variations from word to word, alphabet to alphabet. By making use of the principles of forensic handwriting analysis like consistency, individuality, natural variation, qualitative and quantitative analysis OCR models can be trained.[14]

Interface Development: Currently available software like Tesseract requires in depth knowledge in coding and is time consuming. This reduces the feasibility and utilization by researchers in the field of ayurveda. User-friendly interfaces will be developed to allow scholars and researchers to easily upload manuscript images and train OCR models. The interface will provide a feature for manually uploading a few samples of the confusing characters or hand written alphabets in a particular manuscript thereby training the OCR model specifically capable of transcribing the selected manuscript. The interface guides users through the training process, providing feedback and suggestions to improve model accuracy.

RESULTS

By incorporating machine learning techniques and forensic handwriting analysis for ayurveda manuscripts written in Sanskrit language, a remarkable improvement in the accuracy and efficiency in transcription of the documents in to digital, editable form can be achieved. Research done in English handwritten character recognition provided reliable results which substantiates the proposed approach [15]. Trained OCR models of the English language exhibit significantly higher accuracy in transcribing handwritten manuscripts compared to conventional OCR software.

Scholars will be able to upload the first few pages of the required Sanskrit ayurveda manuscript. After the primary screening, the software will provide the transcribed results of the input. In depth comparison of the accuracy of the characters can be done and the user-friendly interface will enable the scholar in further training the OCR model by uploading samples of the required or even the complete Sanskrit alphabets of that selected manuscript.

DISCUSSIONS

The ability to train customized OCR models by uploading samples of each and every alphabet can be crucial for transcribing Ayurveda manuscripts, especially those written in Sanskrit but using scripts from other languages, such as Kannada,

Malayalam, Nepalese, or Sinhalese. Thus, whichever may be the writing system or language in use, a customized OCR model can recognize and transcribe the content in to Sanskrit with accuracy and ease. This can be beneficial, because many ancient manuscripts, while written in Sanskrit, have been transcribed over the centuries into various regional languages and scripts. The association of Sanskrit alphabets with other available scripts will further contribute in the phonetic conversion of the manuscripts.

Furthermore, user- friendly interface when compared to the currently existing software, website and apps which requires high level of coding and technical expertise, can provide research scholars easier access to the technology. Even though the initial training of the OCR model might require a few hours, the finished model can help reducing the time required for complete transcription and digitization in to editable format from weeks or months to a few hours.

CONCLUSION

The huge amount of unexplored ayurveda manuscripts in India and abroad might hold immense potential in contributing to the advancement of Ayurveda principles and practice. All the available manuscripts have to be explored and undergo the process of manuscriptology for the refinement of the document. By utilizing the artificial intelligence and machine learning technology, OCR software can provide a comprehensive solution for the tedious and time-consuming process of transcribing the manuscript. Nevertheless, OCR software capable of customization which can be trained is required for recognition of the nuances of hand written documents.

This novel approach can improve the accuracy as well as considerably reduce the time required for transcription and thereby significantly evolving the process of manuscriptology. This in turn contribute to new discoveries and helps in bringing the unexplored knowledge in these manuscripts to the limelight of ayurveda.

REFERENCES

1. Namami Gange Program. [Internet]. [cited 2023 Sept 18]. Available from: https://namami.gov.in/

2. Anusha B, Resmi B. Scope of manuscriptology in Ayurveda. Journal of Ayurveda and Integrated Medical Sciences. 2021;6(1):284-9. Available from: https://doi.org/10.21760/jaims.v6i01.1221

3. Anoop AK, Sabu NJ, Bindu KK. A review on manuscriptology – retrieval of ancient knowledge. Int J Ayurveda Pharma Res. 2019 Apr;7(4):39-48. Available from: https://www.ijapr.in/index.php/ijapr/article/view/1190

4. Mathuriya M, Chahar DS, Lal R, Sharma S. A review study of manuscriptology in Ayurveda. *Int J Res Ayur Pharm*. 2020;11(3):45-50.

5. Jagran Josh. List of Ancient Indian Scripts [Internet]. 2023 [cited 2024 Sep 24]. Available from: https://www.jagranjosh.com/general-knowledge/list-of-ancient-indian-scripts-1532423847-1

6. Cambridge University Digital Library. Sanskrit Manuscripts [Internet]. 2024 [cited 2024 Sep 24]. Available from: https://cudl.lib.cam.ac.uk/collections/sanskrit/1

7. J. Memon, M. Sami, R. A. Khan and M. Uddin, Handwritten Optical Character Recognition (OCR): A Comprehensive

Systematic Literature Review (SLR), in IEEE Access, vol. 8, pp. 142642-142668, 2020, doi: 10.1109/ACCESS.2020.3012542.

8. Hegghammer, T. OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *J Comput Soc Sc* 5, 861–882 (2022). https://doi.org/10.1007/s42001-021-00149-1

9. Hussain Al-Aqrabi, Mohammad Sh. Daoud, Fatima B. Shannaq. Handwritten Recognition Techniques: A Comprehensive Review. Symmetry. 2024;16(6):681. doi:10.3390/sym16060681.

10. Kiessling B, Rehbein M, Saadane M, Schulz M. Transkribus – A comprehensive platform for handwritten text recognition and processing. J Digit Stud. 2019;10(1):1-14

11. National Mission for Manuscripts. Major manuscript repositories in India [Internet]. Namami.gov.in. 2024 [cited 2024 Oct 4]. Available from: https://www.namami.gov.in/major-manuscript-repositories-india

12. Khan, A., Sohail, A., Zahoora, U. *et al.* A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev* 53, 5455–5516 (2020). https://doi.org/10.1007/s10462-020-09825-6

13. Jiang J, Poloczek M, Mutny M, Krause A. Hyperparameter optimization: foundations, algorithms, best practices and open challenges. arXiv preprint arXiv:2107.05847v2 [Internet]. 2021 Jul 22 [cited 2024 Aug 22]. Available from: https://ar5iv.labs.arxiv.org/html/2107.05847v2

14. Potle N, Chavan SH, Kekane YH, Tembhurne SU, Pandey N, Dalvi YR**.** Variation in Genuine Handwriting While Writing on an Unusual Surface. *J Forensic Sci Res.* 2023;7:25-33. doi: 10.29328/journal.jfsr.1001046.

15. Zanwar, S.R., Bhosale, Y.H., Bhuyar, D.L. *et al.* English Handwritten Character Recognition Based on Ensembled Machine Learning. *J. Inst. Eng. India Ser. B* 104, 1053–1067 (2023). https://doi.org/10.1007/s40031-023-00917-9

CITE THIS ARTICLE AS

Haritha T J, Sourav K T, Resmi B. Artificial Intelligence powered OCR models for digitizing Ayurveda manuscripts. *J of Ayurveda and Hol Med (JAHM)*. 2024;12(9):24-30

Conflict of interest: None

Source of support: None