

Semantic Analyzer for Sanskrit Scripts

Prof. Bharath Bharadwaj B S¹, Sadaf Sultana², Melvin Dsouza³, Yashaswini S⁴, Darshan Gowda N⁵

¹Assistant Professor, Dept. of Computer Science and Engineering, Maharaja Institute of Technology, Thandavapura.

^{2,3,4,5}Students, Dept. of Computer Science and Engineering, Maharaja Institute of Technology, Thandavapura.

Abstract - The "Semantic Analyzer for Sanskrit Scripts" is a revolutionary system using advanced deep learning technologies to allow modern audiences access to ancient Sanskrit texts. Sanskrit offers philosophical and scientific as well as literary knowledge. However, because it is complex so few experts remain, people battle for understanding it. CNNs are being used in this project in order to achieve accurate recognition. Sanskrit characters are translated into more accessible languages such as Kannada or English. A user-friendly interface lets users easily upload images or text files and the system preprocesses the input for the model's use. Here we have build a model that identifies and translates the characters and gives literal translations with semantic analyses that supply contextual meanings with interpretations. Researchers, students, and enthusiasts worldwide can use this tool for valuable Sanskrit knowledge since it bridges ancient wisdom with contemporary understanding via democratized access.

Key Words: Semantic Analyzer for Sanskrit Scripts, Deep learning, CNN, Upload Image, Preprocess, Literal Translations, Contextual Meanings.

1. INTRODUCTION

This project designed the huge knowledge in ancient Sanskrit texts to unlock. Sanskrit stands as one of the oldest languages in the world, thus it has been a vessel for deep wisdom within various domains, to include philosophy, science, plus literature. The complexity that is in the language has made the comprehension of these texts a challenging endeavor for everyone. Also, a decline in skilled experts has created access problems. For making this rich heritage more accessible to a broader audience, this project seeks to address this issue by employing cutting-edge technology to simplify Sanskrit scripts analysis and interpretation.

A sophisticated deep learning model forms the central component of this system which learns to detect and convert Sanskrit characters into Kannada or English languages. The model employs advanced architectures including Convolutional Neural Networks (CNNs) and Transformers for processing the complex structures present in Sanskrit scripts. The model learns through the training process which requires providing extensive datasets containing labeled Sanskrit characters and their translations to achieve high predictive accuracy. The system produces translations that

maintain both precision and maintain proper context within the output.

The interface of the system provides an accessible platform through which all users can submit images and text files without any prior Sanskrit knowledge. The system begins by pre analyzing the uploaded data to make it suitable for deep learning model evaluation. The model performs character recognition on the input data before generating translations. The system includes semantic analysis functions which deliver contextual meanings to users thus assisting them in understanding texts better. The Semantic Analyzer for Sanskrit Scripts combines advanced technology with user-friendly functions to support researchers and students and language enthusiasts who want to explore the ancient and profound language.

2. PROBLEM STATEMENT AND OBJECTIVE

2.1 Problem Statement

The challenge of accessing ancient Sanskrit texts stems from their linguistic complexity combined with insufficient expert availability. Sanskrit contains an extensive library of philosophical and scientific and literary knowledge yet its complex grammar and script make it difficult to understand. Researchers as well as students and enthusiasts face significant obstacles in their exploration of valuable texts due to both their complexity and the limited number of experts. The traditional translation and analysis methods require expert knowledge and extensive time which leads to the exclusion of the general public from accessing the extensive Sanskrit literature. The limited accessibility to the extensive ancient wisdom means most people cannot explore this valuable knowledge.

The current tools and resources available for Sanskrit translation and analysis fail to meet the necessary standards for proper understanding outdated. Modern translation systems do not possess the needed complexity to decipher the precise meanings of Sanskrit texts. The failure to interpret meaningful nuances in Sanskrit leads to incorrect translations which in turn obstruct text comprehension. The inadequate capabilities of existing systems create an urgent demand for an advanced and intuitive platform that delivers precise Sanskrit translations along with meaningful script interpretations. A modern solution would create wider access

to the vast knowledge stored in Sanskrit by providing better language support.

2.2 Objective

The project aims to build a Semantic Analyzer which uses advanced technologies for translating Sanskrit texts into meaningful interpretations in the modern languages. Through the implementation of deep learning models that use Convolutional Neural Networks (CNNs) alongside Transformer-based architectures the system aims to surpass the conventional translation approaches. The development objective includes producing a tool that converts Sanskrit characters into Kannada and English while maintaining their original cultural and contextual elements.

The second major goal involves making Sanskrit texts more available and easier to use for all types of users. The team wants to create an easy-to-use interface that enables users of different language experience levels to submit Sanskrit text for analysis. The system provides precise translations and semantic analysis to help people better understand Sanskrit literature which makes this ancient knowledge more accessible to everyone while advancing its study and appreciation globally.

3. RELATED WORK

The development of Devanagari handwritten character recognition systems has gained momentum because of deep learning progress and increased dataset sizes. Research teams have made advancements in recognition system accuracy and efficiency by addressing the dual script complexity and joined characters elements.

The initial solutions depended on standard methods that included Histogram of Oriented Gradients (HOG) and Support Vector Machines (SVM) because these techniques encountered challenges when handling Devanagari scripts [1]. Convolutional Neural Networks (CNNs) fundamentally transformed the field because they enabled automatic feature learning from images which led to significant improvement of recognition accuracy.

Researchers have expanded their work through hybrid models which bring together deep learning features with traditional SVM and Random Forest classifiers. The methodology unites the benefits of both systems to achieve better system performance. The combination of multiple models through ensemble methods produces better results and increased reliability.

New possibilities emerged after attention mechanisms and Transformer-based models entered the scene. Through this adaptation of natural language processing methods, models can now detect relevant image elements while tracking distant connections that enhance their performance under limited data conditions.

The development of data augmentation methods alongside transfer learning has mitigated the issue of insufficient training data. Dataset diversity expands through the use of rotation and scaling techniques while pre-trained models through transfer learning speed up development processes and decrease data volume requirements.

Segmentation faces persistent obstacles despite ongoing advancements in the field. The intricate nature of Devanagari scripts creates persistent challenges which researchers attempt to solve through attention-based segmentation algorithms. The research community works on creating models which work efficiently on low- resource devices to guarantee their wider accessibility. Developers now concentrate on producing adaptable systems that users can easily deploy in practical settings.

4. SYSTEM DESIGN

The Semantic Analyzer for Sanskrit Scripts serves to modernize ancient Sanskrit manuscripts through its ability to transform them into digital content that current users can understand. The system follows a detailed and methodical workflow to achieve this goal. The process starts by obtaining digital images of Devanagari manuscripts. The Manuscript Image Database functions as the storage system which contains all the scanned documents scheduled for processing.

Digital images undergo pre-processing once they have been obtained. The important first step in this stage requires image improvement alongside noise elimination and slant and skew correction to produce documents which are suitable for further analysis. After pre- processing, the document images move to the Segmentation module. The system operates within this module to break down the text content into three-part segments that represent lines and words and single characters while recording the spatial details for each segment.

The system proceeds to Feature Extraction following the segmentation process which enables it to detect and extract special characteristics of each character. The features extracted in this stage serve as the foundation for the Classification step where the trained model identifies each character class. To validate correct character identification the system matches classified characters against a Character Classes Database.

The system incorporates an Error Detection and Correction module to maintain the original content integrity of the recognized text. The module operates by detecting classification mistakes and then performing required adjustments. The system applies Post Processing to the processed text during which it transforms the content into a well-organized format that improves its readability for users.

Semantic Analyzer stands out because it can transform Sanskrit words into English and Kannada languages. After text recognition and classification the system initializes translation algorithms to change Sanskrit characters into target languages. Through this translation method the knowledge from Sanskrit manuscripts becomes available to all readers regardless of their language background.

The process for designing the Semantic Analyzer for Sanskrit Scripts focuses on extensive detailing to convert Devanagari manuscript handwriting into digital text. The system achieves high accuracy and usability through its implementation of sophisticated image processing and feature extraction alongside segmentation, classification, error correction and post-processing techniques. The translation functionality improves the availability of ancient Sanskrit knowledge to modern users by producing content in different languages.

After pre-processing, the system goes to Segmentation. This stage is divided into small processes: line segmentation, word segmentation and character segmentation. Characters are discriminated with state-of-the-art algorithms, such as connected component analysis and contour detection, and here the somewhat complex operation which can separate characters even if connected or added with complex details. When the text is partitioned, the system proceeds to Feature Extraction. This is an important step, in which the features that uniquely belong to each author are identified and extracted. You can use Histogram of Oriented Gradients and Convolutional Neural Networks. CNNs, especially, are very powerful for this task given their capacity to learn the hierarchical features automatically from the raw images. The features are then used for classification.

The Classification phase employs deep learning models, which are mostly CNNs and Long Short-Term Memory (LSTM) networks, to identify and classify each character [1]. CNNs are particularly good at extracting spatial hierarchies from images, and LSTM networks are well-suited to handle sequential data, thus making them good at recognizing characters in context. The models are trained on a vast dataset of labeled Devanagari characters so that the models are highly accurate and resilient.

To enable the translation of Sanskrit text into other languages, the system has an API for Translation. The API uses machine translation algorithms, including those based on Transformer models, to translate the recognized Sanskrit text into languages such as English and Kannada. The API is capable of translating the nuances and complexities of Sanskrit and providing accurate and meaningful translations [18]. here we have used google/flan-t5-base model to translate from Sanskrit to English and facebook/nllb-200-distilled-600M model for English to Kannada translation.

The system also comprises an Error Detection and Correction module. This module applies algorithms to detect and correct any mistakes in the classification and translation process. Methods like confidence scoring and contextual analysis are utilized to provide reliability to the final output. Lastly, the processed text is subjected to Post-processing, during which it is formatted and polished for readability. This involves activities like text alignment, punctuation addition, and language translation. The system makes sure that the final output is readable and understandable to users.

In short, the Semantic Analyzer for Sanskrit Scripts combines deep learning algorithms (CNNs and LSTMs), high-level image processing algorithms, and a translation API to develop an exhaustive tool for digitizing and deciphering ancient Sanskrit manuscripts. By utilizing these high-tech technologies, algorithms, and methodologies, the system attains high accuracy, efficiency, and usability to provide wider access to the rich legacy of Sanskrit.

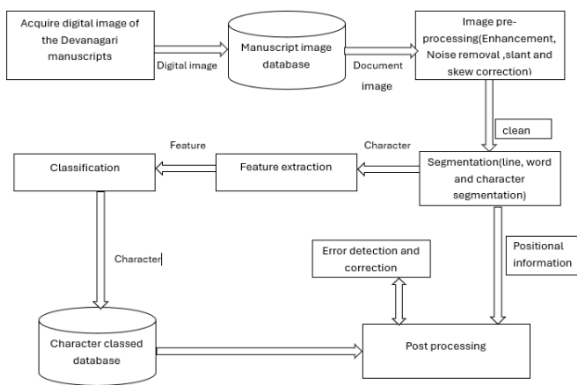


Fig -1: System Architecture for Semantic Analyzer for Sanskrit Scripts

5. METHODOLOGY

The Semantic Analyser for Sanskrit is an advanced system which uses a mix of high technology and algorithms to convert the centuries old Sanskrit manuscripts into digitised text format and also translate. Project As indicated by the systematic approach, the project is divided into four phases, beginning with Data Acquisition where the manuscripts are digitized process whereby scanners or cameras are used to capture the documents. Such images are maintained in the form of a centralised Manuscript Image Database that serves as the central data source of the overall workflow.

The next crucial part is the "Image Pre- processing". Namely the quality of the digital images will be improved when they are optimal for the analysis. "Techniques like reducing noise, sharpening images and geometric distortions, such as compensating for the slant and skew, which are also applied. Sophisticated pre- processing techniques guarantee a clear text which is free from any deformation that may interfere with the correct processing.

6. IMPLEMENTATION

The Semantic Analyzer for Sanskrit Scripts is a high-tech system that aims to digitize and translate ancient Sanskrit manuscripts into usable digital forms. The process starts with Data Acquisition, where high-resolution images of the manuscripts are taken and stored in a centralized Manuscript Image Database. This is then followed by Image Pre-processing, where methods such as Gaussian Blur, Median Filtering, Histogram Equalization, and CLAHE improve image quality, while Affine and Perspective Transformations rectify geometric distortions to provide clean and standardized images [19].

The system then moves on to Segmentation, using sophisticated algorithms like Projection Profile Methods, Connected Component Analysis, and Contour Detection to separate the text into lines, words, and individual characters, even if they are connected or have intricate modifiers. Feature Extraction comes next, using Histogram of Oriented Gradients (HOG) and Convolutional Neural Networks (CNNs) to detect the shape and structure of characters [20]. These aspects are instrumental in the Classification phase, in which deep learning models such as CNNs and Long Short-Term Memory (LSTM) networks identify and classify every character with high accuracy. The identified text is next translated to English and Kannada using an API driven by Transformer models, which take care of Sanskrit's nuances [11]. Subsequently, the translated text receives Post-processing to make it more readable, viz., Text Alignment, Insertion of Punctuation, and Language-Specific formatting.

6.1 Key Performance Indicators

The system is top in many aspects: recognizes characters accurately, handles large amounts of text rapidly with low latency, and is very scalable, with a vast variety of document sizes and complexity. Ensemble methods and error detection mechanisms increase its resilience, making it capable of coping with handwriting and document quality variations.

6.2 Usability and User Experience

The system has a user-friendly interface that makes upload and analysis of manuscripts easy. Users are provided with real time feedback on the digitization and translation process, and the output is displayed prominently, with the facility to download in multiple formats. Facilities such as User Authentication and Data Management enable users to save and retrieve their work, making it ideal for scholars and researchers. The interface is simple and user-friendly, accommodating users with different degrees of familiarity with Sanskrit and computer tools.

6.3 Limitations and Challenges

In spite of its strengths, the system has limitations. The joined nature of the Devanagari script and its modifiers makes

it challenging to achieve accurate segmentation, and certain errors can be produced in low-quality documents. Large and varied datasets are essential to the accuracy of deep learning models, and the lack of labeled data for some languages and scripts is still an issue. Also, the utilization of sophisticated models demands large computational resources, which can impact processing times for those with low-end hardware.

7. CONCLUSION

The Semantic Analyzer for Sanskrit Scripts is a breakthrough in the technology of optical character recognition (OCR) and machine translation with a complete solution for digitization and translation of ancient Sanskrit manuscripts. With the incorporation of sophisticated technologies like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models, the system boasts high accuracy in character recognition and translation, serving as a powerful tool for scholars, researchers, and enthusiasts [10]. The use of advanced image processing methods ensures that the system can effectively deal with the intricacies of the Devanagari script, including joined characters and modifiers, with robustness against different document qualities. The ease of the system's use and its live feedback processes provide increased usability for users of any level of exposure to digital systems and to Sanskrit. Notwithstanding limitations from the demands for large, multicultural datasets and computationally intensive systems for deep models, the Semantic Analyzer is evidence of technology as a powerful facilitator for storing and spreading information from antiquity. With the system improving, it is not only going to make Sanskrit manuscripts more readily available but also help understand and appreciate this rich cultural and linguistic heritage to a greater extent.

8. FUTURE ENHANCEMENT

The Semantic Analyzer for Sanskrit Scripts has established a solid foundation for digitizing and translating ancient Sanskrit manuscripts, but there are a number of areas for future enhancement to further enhance its capabilities and usability. One area of focus for improvement is segmentation of sophisticated Devanagari character. Although the system today uses more advanced algorithms, it is still possible to improve these methods further in order to handle more complex scripts and handwriting variation better. Adopting more advanced segmentation techniques, including deep learning segmentation models, has the potential to improve accuracy considerably and minimize error.

Another significant improvement is the increase in the training dataset. The system's performance currently relies significantly on the presence of labeled data. By developing a larger and more diverse dataset containing a greater variety of scripts, handwriting styles, and languages, the system can enhance its capacity to recognize and translate a greater

variety of Sanskrit texts. This may involve working with institutions and experts to collect and annotate more data.

Technically, the capability of translation capabilities can be made even more efficient by incorporating stronger natural language processing (NLP) methods for further enhancing translation accuracy and flow. That could involve context-aware translation algorithms that take the overall context of a paragraph or sentence into consideration, instead of translating word-for-word. Also, extending the scope of target languages past English and Kannada would make the system more general and beneficial for a worldwide group of people. The system might also be improved by more user features. For example, the inclusion of a user feedback loop in which users can annotate or correct translations would enable the system to learn and improve over time. Adding collaborative features, like the ability to share and discuss translations among a community of users, would also make the system more useful for research and educational applications.

Lastly, the system needs to be optimized for performance and scalability. With the number of manuscripts and data complexity on the rise, it will be essential to ensure that the system can efficiently manage increasing workloads. This could include the use of cloud-based infrastructure for storage and processing, and better algorithms to minimize the computational load.

REFERENCES

- [1] AI Based OCR for Ancient Indian Text,2025.
- [2] Chetan Sharma, Shamneesh Sharma, Advancements in Handwritten Devanagari Character Recognition, 2024.
- [3] Sandhya Arora, Latesh Malik, Sonakshi Goyal, Devanagari Character Recognition: A Comprehensive Literature Review, 2024.
- [4] M. P. Ayyoob, P. Muhamed Ilyas, Stroke-Based Data Augmentation for Enhancing OCR of Ancient Handwritten Scripts,2024.
- [5] Dr. K. Abdul Rasheed, From Grammar to Algorithms: The Digital Transformation of Sanskrit Studies,2024.
- [6] Anchal Chand, Piyush Agarwal, Sachin Sharma Real-Time Retrieving Vedic Sanskrit Text into Multi- Lingual Text and Audio,2023.
- [7] Shraddha V. Shelke, Dr. S.P. Ugale, Combining Multiple Feature Extraction and Classification Methods to Study handwritten scripts,2023.
- [8] Arpit Sharma, Mithun B N, Deep Learning Character Recognition of Handwritten Devanagari Script,2023.
- [9] Khushi Sinha, Eshan Marwah, Richa Gupta, Deep Learning Based Enhanced Handwritten Devanagari Character Recognition using Image Augmentation,2023
- [10] Kavita Bhosle, Evaluation of Deep Learning CNN Model for Recognition of Devanagari Digit, 2023.
- [11] Sandeep D. Pande, Pramod P. Jadhav, Rahul Joshi, Digitization of Handwritten Devanagari Text using CNN Transfer Learning,2022.
- [12] Ch.Venkata Sasi Deepthi, A. Seenu, A Systematic Review on OCRs for Indic Documents & Manuscripts, 2022.
- [13] Vina M. Lomte, Dharmal D. Doye, Handwritten Vedic Sanskrit Text Recognition Using Deep Learning,2022.
- [14] Pragati Hirugade, Nidhi Suryavanshi, Radhika Bhagwat, A Survey on Optical Character Recognition for Handwritten Devanagari Script Using Deep Learning,2022.
- [15] Bhavesh Kataria, Optical Character Recognition of Sanskrit Manuscripts using Convolution Neural Networks,2022.
- [16] Ayush Maheshwari, Nikhil Singh, Benchmark and Dataset for Post-OCR Text Correction in Sanskrit,2022.
- [17] Kulkarni, Pandit, Kharate, Tikkal, Chaware, Proposed Design to Recognize Ancient Sanskrit Manuscript with Translation Using Machine,2022.
- [18] S.D. Pande, P.P. Jadhav, R. Joshi, Digitization of handwritten Devanagari text using CNN transfer learning,2022.
- [19] Pankaj Tukaram Bhise, Vina Lomte, Darshan Derle, Sanskrit Text Recognition using Machine Learning,2022.
- [20] Ranadeep Dey, Pranav Gawade, Ria Sigtia, Shrushti Naikare, Atharva Gadre, Diptee Chikmurge, A Comparative Study of Handwritten Devanagari Script Character Recognition Techniques,2022.