Project Proposal

Level 4

Context-Aware Damage Detection and Text Restoration in Sinhala Handwritten Documents

Group Name: Reformers

Faculty of Information Technology

University of Moratuwa

2024

Project Proposal

Level 4

Context-Aware Damage Detection and Text Restoration in Sinhala Handwritten Documents

Group Name: Reformers

Group Members

Index Number	Name
204009V	**Athukorala D.A.Y.S**
204044X	Disara M.P.A
204065L	Gunathilaka M.D.K.L
204190N	Sankalpana B.L.P

Supervisor: Dr. L. Ranathunga

Faculty of Information Technology

University of Moratuwa

2024

Table of Content

1. Introduction

Sri Lanka has a rich history that dates back thousands of years and continues to this day in the 21st century. The ancient ancestors dealt with several languages in advancing the culture, including Pali and Sanskrit, which is believed to be the language that gave rise to Sinhala. With most of its letters being rounded, Sinhala writing differs greatly from most other languages used today. Sinhala is not for the faint-hearted only a very motivated individual can learn to read and write it. A significant portion of our historical knowledge is derived from numerous old books, pamphlets, manuscripts, and chronicles authored by our ancestors. A large number of them are accessible in Sinhala. Unfortunately, insect attacks, fading handwritten ink, wear and tear, and ink blotches cause valuable documents to be destroyed and content to be lost. It might be difficult to reconstruct handwritten Sinhala documents after they have been destroyed while maintaining their value. The project that has been suggested could be presented as a solution in this specific situation.

The project's goal is to assist in resolving some of the problems pertaining to the damage restoration of handwritten Sinhala documents, which are crucial to the preservation of historical records, legal documents, and scholarly materials. This includes estimating the missing text, narrowing the damaged areas, and returning the document to its initial state through the application of Deep Learning(DL) and Machine Learning (ML) techniques. Using a customized algorithm to detect the damaged letters and words, this project extends towards the restoration of missing content with the help of image processing and reconstruction of missing words and phrases using Natural Language Processing(NLP).

It's crucial to restore a handwritten document that has been damaged for several reasons. It aids in the preservation of historical and cultural records, which are essential for helping present and future generations comprehend and learn from the past. These documents can then be made available online for global access, study, and learning after they have been restored. This encourages the exchange of knowledge, making these priceless resources available to a far wider audience.

2. Background & Motivation

The motivation stems from the need to restore Sinhala handwritten documents, which contain culturally significant information relevant to many aspects of ancient Sri Lanka. As Sri Lanka is the only country that uses Sinhala as a native language, it is obvious that it still has very little research done. The lack of technology and implementation to work with Sinhala letters stressed the importance of this initiative even more.

Regarding the restoration of damaged Sinhala written text using image processing, the necessity to address the limitations they have faced was felt important. In that particular research, a green colored background was placed to detect the damaged part of a particular letter, which cannot be considered a perfect approach for damage detection. Moreover, the studies have only addressed the restoration of a part of the letter by linking broken edges using techniques. While this method achieved a reasonable level of accuracy it was primarily focused on reconstructing parts of single letters and could not restore entire words or phrases. Additionally, the accuracy of these techniques decreased when handling complex characters or 90-degree angled letters. Moreover, this research is entirely image processing-based and this approach, while useful for reconstructing small segments, cannot infer missing textual content in complex or large-scale damages where significant portions of text are missing. It also fails to incorporate contextual understanding, leading to errors when reconstructing complex letters or phrases. [1],[2]

Apart from that, handwritten documents often tend to deteriorate over time due to environmental factors, aging, and mishandling. This project seeks to develop a custom damage detection and restoration pipeline tailored for handwritten documents, improving the accuracy and quality of restoration. By developing an automated damage detection and restoration approach, this project aims to improve the process of document preservation and make valuable information more accessible.

3. Problem in Brief

The project concerns the development of a system able to identify the location and restore damaged portions within Sinhala handwritten documents. Handwritten documents can be destroyed by ink spills, tears, bug damage, and fading due to aging, environmental factors, and improper handling. Due to this fact, much valuable information has been lost, which is very hard to maintain about the cultural, historical, and legal values of the documents.

The problem mainly decomposes into two parts: finding the location of the damaged areas in this document correctly. In other words, it is the detection of edges of the damaged spots-for example, small torn-out pieces, lack of ink, or faded areas-without necessarily needing human intervention. The task requires a robust identification system for damages that should handle multiple types of damages while recognizing damaged portions from undamaged portions of the document.

Once the location of the damages has been found, the next task is to determine what is missing in these places. That is, by observing the document's layout, such as the distance between the letters, the length of the words, and the space between paragraphs, the system can guess what text might be missing. The system needs to find out whether the damage hurts just one letter, part of a word, or even several phrases. The important thing is to make the missing content cohere appropriately with its surrounding text, both semantically and syntactically.

This project is related to text restoration, which has been done using machine learning and various NLP techniques. The models used in this paper, such as transformers, use the surrounding context of the text to fill in the missing words or phrases by merely guessing. The solution ensures that the restored text fits within the document making sure that it delivers the proper meaning and that it is clear and accurate.

This method is different from regular image processing techniques. It aims to provide a smarter system that understands the context for fixing damaged handwritten documents, making sure they are saved for future use.

4. Aim & Objectives

4.1 Aim

This project strives to create a method for finding and fixing damaged sections in handwritten writings. It uses special object detection models, picture processing tricks, and transformer-based text rebuilding.

4.2 Objectives

1. Look into the problem and understand existing ways of finding damage.
2. Create a customized model to spot the boundaries of damage in texts.
3. Use techniques to determine how much text is missing using word count and distance between paragraphs.
4. Employ image processing to identify unclear or blurred parts of words.
5. Completion of missing words or phrases by transformer models, like GPT and BERT.
6. Identify how the solution will work for different kinds of handwritten notes.

5. Proposed Solution

To solve the above issue, it is necessary to identify any damage in handwritten Sinhala documents and reconstruct them by adding the missing text. This solution will be applied at the following stages:

1. Module 1 - Detect Damage Areas
2. Module 2 - Determine Missing Parts
3. Module 3 - Reconstruct Word Parts/Blur (Image Processing)
4. Module 4 - Reconstruct Phrases (Transformers)

To achieve this solution, Machine Learning techniques for damage detection, Image Processing for visual quality restoration, and Natural Language Processing (NLP) for text reconstruction are integrated.

5.1 Module 1 - Detect Damage Areas

Detecting damaged areas is the most initial part of the project. This module is responsible for developing a model that can figure out what regions are damaged in handwritten documents. Hence, at this point, it will segment the document into subsections and recognize the region that has been damaged by ink smudge, insects, holes, tearing, or faded text. [5]

- Model Design

  The model is proposed to recognize damage regarding handwritten documents while identifying the perimeters of ink blotches, paper tears, and other forms of damage.

- Training Data

  Handwritten documents with pre-marked annotated regions indicating damage.

5.2 Module 2 - Determine Missing Parts

- This module concerns how the document is set up to find what parts of the text are missing. The number of words, phrases, or parts of a word missing are detected by using the patterns found in the document.[8]
- Document Layout Analysis

  The document layout analysis involves using OCR (Optical Character Recognition) techniques to study how the text is arranged. This model seeks for the average size of different types of words, the space between letters, and the breaks between paragraphs in the document. It then guesses any missing characters, words, or lines after examining these features.[4]

- Knowledge Acquisition from Similar Texts

The system enhances the prediction of missing parts by studying similar texts of handwriting. This will help the model to learn typical patterns and uses of words, which could help it in making predictions.

5.3 Module 3 - Reconstruct Word Parts / Blur (Image Processing):

In the third module, the reconstruction of damaged portions like missing parts of words or blurred sections will be done using Image Processing techniques.

- Super-Resolution

  The super-resolution algorithms enhance the quality and clarity of blurred text to make it readable. By training the model on low-resolution-high-resolution image pairs of handwriting images, the model can learn to sharpen blurred text.[3]

- Preprocessing and Noise Reduction

  Before applying reconstruction techniques, preprocessing methods like noise reduction can be used to clean up the document image and improve the performance of the restoration process.

- Inpainting for reconstructing word parts

  Develop a custom inpainting solution tailored specifically for the reconstruction of missing word parts. This technique will be engineered to rebuild torn or missing fragments of words by intelligently filling in the missing strokes in the handwriting through analysis of the surrounding strokes and natural flow of the handwriting document.[7]

5.4 Module 4 - Reconstruct Phrases (Transformers)

The final module will be used to make predictions of missing phrases or even larger parts of the text using transformer-based models, by taking into consideration the context from the surrounding text.[6]

- Contextual Understanding

  By training a transformer model like GPT or BERT on handwritten document datasets, the model will learn to fill in missing phrases based on what it learned of the surrounding words of a broken part of a document with datasets of handwritten documents.

- Fine-Tuning for Handwritten Text

Although transformers are inherently trained on typed texts, the model would need fine-tuning with handwritten documents to handle the style variations of different handwritings. This will thus enable it to create text that closely resembles the original document.

- Matching Text and Style

  The output of the transformer needs to be semantically correct, but it also needs to match the style and tone of the surrounding text. Here, post-processing steps can be taken to ensure that the reconstructed text matches the writing style of the original.

The above-proposed solution advances the previous studies limitation such as,

- Detecting damaged areas directly from the document image, independent of external aids like colored backgrounds.[1][2]
- Instead of focusing only on reconstructing parts of individual characters, intelligently restore entire words or phrases based on the context of the surrounding text.[1][2]
- Integrates NLP techniques (Natural Language Processing) for text reconstruction, without solely focusing on image processing.[1][2]

6. Resource Requirements

- Software Requirements:
    - Python with libraries such as TensorFlow, PyTorch, OpenCV, and Transformers(Hugging Face)
    - Jupyter Notebooks, and Google Colab for development and experimentation.
    - Git/GitHub: For collaboration, version control, and tracking changes in the project.
    - LabelImg: For labeling Dataset
    - Tesseract OCR: For recognizing characters from images
- Hardware Requirements:
    - High-performance GPU for training deep learning models.
    - Standard computing resources for dataset preparation and model evaluation.
- Datasets:
    - A collection of handwritten document images with various types of damage for training and testing.
    - Manually labeled data for damage detection and text reconstruction.

7. Reference

[1] H. K. I. L. Madhuwanthi and L. Ranathunga, “Reconstruct Damaged Sinhala Handwritten Characters by Guided Edge Linking Approach,” Feb. 2023, doi: https://doi.org/10.1109/icarc57651.2023.10145633.

[2] L. B. V. Lakshitha, H. K. I. L. Madhuwanthi, and L. Ranathunga, “Damaged Sinhala Handwritten Character Location-Identification using Neighbour Mapping,” Feb. 2023, doi: https://doi.org/10.1109/icarc57651.2023.10145729.

[3] N. Sandhya and R. Krishnan, “Broken kannada character recognition — A neural network based approach,” vol. 9, pp. 2047–2050, Mar. 2016, doi: https://doi.org/10.1109/iceeot.2016.7755047.

[4] P. Bannigidad and C. Gudada, “Restoration of degraded Kannada handwritten paper inscriptions (Hastaprati) using image enhancement techniques,” *IEEE Xplore*, Jan. 01, 2017. https://ieeexplore.ieee.org/abstract/document/8117697 (accessed Nov. 27, 2021).

[5] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning,” *MIT Press*, Nov. 18, 2016. https://mitpress.mit.edu/9780262035613/deep-learning/

[6] A. Vaswani *et al.*, “Attention Is All You Need,” *arXiv*, Jun. 12, 2017. https://arxiv.org/abs/1706.03762

[7] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image Inpainting for Irregular Holes Using Partial Convolutions,” *arXiv:1804.07723 [cs]*, Dec. 2018, Available: https://arxiv.org/abs/1804.07723

[8] R. C. Gonzalez, R. E. Woods, and B. R. Masters, “Digital Image Processing, Third Edition,” *Journal of Biomedical Optics*, vol. 14, no. 2, p. 029901, 2009, doi: https://doi.org/10.1117/1.3115362.

Action Plan link -

Group 37- Reformers Action plan