INFORMATICS INSTITUTE OF TECHNOLOGY

In collaboration with

UNIVERSITY OF WESTMINSTER

Recognizing Tamil Character from Noisy Background in Historical Documents Using GAN (Generative Adversarial Networks)

A dissertation by
Mr. Haniman Ananthamoorthy
2018556 (w1742243)

Supervised by
Mr. Ragu Sivaraman.

Submitted in partial fulfilment of the requirements for the BEng in Software Engineering degree at the University of Westminster.

May 2022

ABSTRACT

Archaeologist department it's a very famous and vast area to dive through the past events and make a summary for people history. And they are mainly focusing on finding the root of language, religion, and for the specific region of people. It will bring this specific people's lifestyle and their root of beginning. There are many researchers contributed towards these Archaeologists department and there is much software which makes easier their work, in this research the gap between literature which cannot be renovated, which is hard to renovate who are in different regions of the world. The research almost demonstrates creating a platform to renovate the inscriptions, manuscript, literature books, and any other damaged documents to a digital version in a human-readable way. Designing and implementing GAN model classifier using specific region language datasets driving the model to predict text, and the specific words. The main component is to get the damaged datasets and predict those words to form a meaningful sentence.

Developing a word predicting classifier which is using an image that contains several tasks and different levels, this task has been achieved using the GAN model, and some other machine learning libraries. The biggest task is to collect data sets for each word in a different format and a different style .They implemented a system detecting noisy letters to predict the correct format of the letter with more accuracy level the model will predict the entire word, and because of Deep Learning accuracy and responses are quick .And addition of this feature there will be a blockchain system to save the inscriptions data, manuscript in clusters way for the future generations to ensure that trustworthy of those documents.

Keywords: DE-GAN, Image processing, Optical character recognition, Generative Adversarial Networks, Historical Documents, Ancient Tamil Handwritten Characters