INTERNSHIP REPORT

Generating Synthetic Word Images for training Indian Scripts Scene Text OCR

Submitted by:

RAUNAK NAG (201B374)

Under the guidance of
Prof CV Jawahar
International Institute of Information Technology
Hyderabad, Telangana, India
Jun 12 – Aug 7

Submitted in partial fulfillment of the Degree of Bachelor of Technology

Department of Computer Science & Engineering

2020-2024

JAYPEE UNIVERSITY OF ENGINEERING & TECHNOLOGY,

A-B ROAD, RAGHOGARH, DT. GUNA - 473226, M.P., INDIA

DECLARATION

We hereby declare that the internship work reported in 7th semester entitled "Generating Synthetic Word Images for training Indian Scripts Scene Text OCR", in partial fulfillment for the award of the degree of B.Tech (CSE) submitted at Jaypee University of Engineering and Technology, Guna, as per the best of our knowledge and belief there is no infringement of intellectual property rights and copyright. In case of any violation, we will solely beresponsible.

Raunak Nag (201B374)

Date: 12/08/23

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my guide, Prof. C.V. Jawahar and supervisor Dr. Ajoy Mondal at IIIT Hyderabad for their wonderful guidance, constant help and constant support at each and every step of the project during my entire internship journey. Also, I am very thankful to all of my friends and colleagues whom I have interacted in the college in this duration of 2 months.

Raunak Nag (201B374)

Date: 12/08/23

EXECUTIVE SUMMARY

Purpose: Generating synthetic word images in Bengali Language for training Optical Character Recognition (OCR) systems specifically designed for Indian scripts.

Work Done:
Dataset Collection: Gather a diverse set of Indian script characters and glyphs.

Font Variation: Select a range of fonts that accurately represent the typographical diversity found in printed Indian text. This includes various serif, sans-serif, and decorative fonts.

Style Variation: Introduce variations in font size, weight, style (italic, bold), and alignment to simulate real-world text variability.

Background and Noise Addition: Integrate synthetic backgrounds and noise to mimic the visual complexity of scene text captured from natural images.

Data Augmentation: Implement data augmentation techniques like rotation, perspective distortion etc., to increase the diversity of the generated dataset.

Learning Outcome:
Technical Skills: Gained proficiency in various technical areas such as image synthesis, typography, character arrangement, font manipulation, data augmentation, and image processing.

Data Augmentation Strategies: Learned to diversify datasets through augmentation techniques like rotation, scaling, and perspective distortion, applicable in various machine learning and data analysis tasks.

Table of Contents

CHAPTER-1

INTRODUCTION

1.1 About

International Institute of Information Technology Hyderabad (IIITH) is a higher-education institute deemed-to-be-university, founded as a non-profit public private partnership located in Telangana, India. It is the first IIIT in India under this model. Over the years, the institute has evolved strong research programmes in various areas, with an emphasis on technology and applied research for industry and society. The institute facilitates interdisciplinary research and a seamless flow of knowledge. Several world-renowned centres of excellence are part of IIITH's research portfolio. It has established various joint collaboration and co- innovation models with an industry outreach spanning significant national and multinational companies. Its innovative curriculum allows students the flexibility of selecting their courses and projects.

1.2 Project Overview

A dataset of synthetically generated word images specifically but not only limited to different fonts of Bengali language to train Optical Character Recognition (OCR) systems for various Indian Scripts. The aim is to address the challenges of limited labelled data by generating realistic synthetic word images that represent cases of real-word data.

CHAPTER-2

LITERARTURE SURVEY

2.1 EXISTING SYSTEM

In the existing system, training Optical Character Recognition (OCR) systems for Indian scripts faces challenges due to limited labeled data. The available datasets might not cover the diverse range of fonts, styles, and layouts found in real-world text scenarios. This leads to OCR models that might struggle to accurately recognize Indian script text in various contexts, impacting their performance and applicability.

2.2 PROPOSED SYSTEM

The proposed system aims to address the limitations of the existing approach by leveraging synthetic word images for training OCR systems tailored to Indian scripts. This involves generating a diverse dataset of synthetic word images that mimic real-world text scenarios. The synthetic dataset is created by combining characters, matras, vowels, and ligatures in various fonts, styles, and layouts. The generated images are augmented with variations like font size, weight, and background noise. These synthetic images are then integrated with real data to train OCR models.

CHAPTER-3

SYSTEM & DESIGN ANALYSIS

3.1 Requirement Specification

3.1.1 PYTHON

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

3.1.2 COMMAND LINE

A command-line interface (CLI) is a means of interacting with a device or computer program with commands from a user or client, and responses from the device or program, in the form of lines of text.

3.1.3 IMAGEMAGICK

ImageMagick is a free, open-source software suite, used for editing and manipulating digital images. It can be used to create, edit, compose, or convert bitmap images, and supports a wide range of file formats, including JPEG, PNG, GIF, TIFF, and PDF.

ImageMagick includes a command-line interface for executing complex image processing tasks, as well as APIs for integrating its features into software applications. It is written in C and can be used on a variety of operating systems, including Linux, Windows, and macOS.

3.1.4 PANGOCAIRO

Library for laying out and rendering of text, with an emphasis on internationalization

3.1.5 PANGO

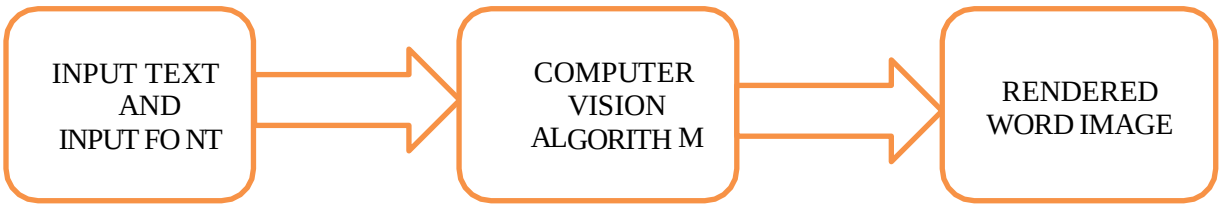
Pango is a library for layout and rendering of text, with an emphasis on internationalization. Pango can be used anywhere that text layout is needed. Pango forms the core of text and font handling for GTK.

3.1.6 CAIRO

The Cairo library is a vector graphics library with a powerful rendering model. It has such features as anti-aliased primitives, alpha-compositing, and gradients.

Multiple backends for Cairo are available, to allow rendering to images, to PDF files, and to the screen on X and on other windowing system.

3.2 Block Diagram of the Project



Explanation:

The whole internship project work can be mainly divided into seven phases-

1) Data Preparation- The project uses a collection of fonts compiled for Bengali language. Additionally, a set of images, preferably natural scene images, is needed to serve as background for the rendered word images. Our project uses the Places365 dataset for this purpose. It involves a python script that requires imagemagick, pango, cairo, pangocairo libraries to be installed on the linux machine and invokes bash commands from the command line.

2) Reading a specific language vocabulary file- In our case, we have chosen Bengali as our desired language. The project starts by reading a vocabulary file containing a list of words that need to be rendered as synthetic images. The script then reads each word from the file and then proceeds to render the word image. L SEP

3) Choosing suitable rendering parameters- Here, for each word, the script randomly selects various rendering parameters like font name, font stretch, font color, font style, font size, foreground (text) color and background color along with perspective distortion.

4) Rendering the text word as an image- Using the selected rendering parameters, the script invokes the Pango library to render the word as an image by choosing a random font from a list of installed fonts each time the program runs. The rendered image is then saved as a JPG file.

5) Background Blending- The script randomly selects the background should be a natural scene image blended with the text layer or a uniform color. We have chosen the former case where the script takes a random crop from the Places dataset and blends it with the text layer and kept the latter one optional.

6) Image Post-Processing- The script applies optional image post- processing, including erosion, dilation, gaussian blur, median blur, salt and pepper noise and gaussian noise. These post- processing operations add variations to the synthetic images, making them more diverse and suitable for training OCR models.

7) Saving the Final Image- The final synthetic word image, along with its rendering parameters and other metadata, is saved in a specified output directory and all the values of the rendering parameters are written to a detailed annotation csv file for Bengali language.

CHAPTER-4

RESULTS/OUTPUTS

Figure: Sample images of rendered Bengali text word images from our synthetically generated dataset

CHAPTER-5

CONCLUSION

- I learnt the proper use of command line while setting up my project and also used different terminal commands for various file-related operations and functions.
- I understood the importance of writing clean code with proper variable and function names along with multi-line comments for proper user readability and user clarity.
- I learnt various ways of debugging code for issues and errors and to manage them properly in a research-oriented way.
- I learnt the importance of synthetic data and understood their generation process.
- I came to know about the importance of synthetic data and their role in training OCR in Indian scripts.

CHAPTER-6

REFERENCES

- https://en.wikipedia.org/wiki/Scene_text
- https://imagemagick.org/index.php
- https://docs.gtk.org/PangoCairo/pango_cairo.html
- https://cvit.iiit.ac.in/research/projects/cvit-projects/iiit-ilst