**GUIDELINES FOR B.TECH. 6TH SEM PROJECT REPORT WRITING, DEPT. OF IT-GU**

****

**Department of Information Technology**

**Gauhati University**

**Assam :: India**

**Guidelines for preparation of the report**

**Report Binding**

The report should be spiral binding

**Paper Size**

The thesis should be typed in A4 size paper in 1.5 space with font size of 12 point.

**Margin**

Margin of at least 3 cm shall be left on all sides.

**Number of copies**

A minimum number of copies to be 01 (one)

**Before producing final copies of a report for submission, the candidate should ensure that all the spellings, punctuation, vocabulary and grammar are correctly used.**

**Organization of the report**

The report may be organized in the following manner

* Title page
* Declaration by the Candidate
* Certificate from guide (in prescribed format)
* Acknowledgments
* Abstract
* Table of contents
* List of tables, figures
* The main chapters may be divided into several parts
  1. Introduction
  2. Background/ Review of literature
  3. Methodology /System Architecture
  4. Propose Solution for your problem (If any)/Implementation
  5. Results and Discussions (if Any)
  6. Conclusion and Future Works
* References
* Appendices (wherever necessary)

**ASSAMESE LANGUAGE OPTICAL CHARACTER RECOGnition MODEL USING PYTORCH**

**REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**By**

**Rishav Bora**

**Roll Number: 220103011**

**Rishikesh Verma**

**Roll Number: 220103004**

**UNDER THE GUIDANCE**

**OF**

**Kishor Sir**

**Assistant Professor**

****

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**GAUHATI UNIVERSITY**

**GUWAHATI, INDIA**

**JUNE –2025**



GAUHATI UNIVERSITY

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**Gopinath Bordoloi Nagar, Jalukbari Guwahati-781014**

**DECLARATION**

I “Rishav Bora” and “Rishikesh Verma”, bearing roll No “220103011” and “220103004”, a B.Tech. student of the department of Information Technology, Gauhati University hereby declares that I have compiled this report reflecting all my works during the semester long full time project as part of my BTech curriculum.

I declare that I have included the descriptions etc. of my project work, and nothing has been copied/replicated from other’s work. The facts, figures, analysis, results, claims etc. depicted in my thesis are all related to my full time project work.

I also declare that the same report or any substantial portion of this report has not been submitted anywhere else as part of any requirements for any degree/diploma etc.

Rishav Bora

Branch: Computer Science & Engineering

Date: 30-05-2025

Rishikesh Verma

Branch: Computer Science & Engineering

Date:30-05-2025



GAUHATI UNIVERSITY

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**Gopinath Bordoloi Nagar, Jalukbari Guwahati-781014**

Date: 30-05-2025

**CERTIFICATE**

This is to certify that “Rishav Bora” and “Rishikesh Verma” bearing Roll No: “220103011” and “220103004” has carried out the project work “Assamese Language Optical Character Recognition Model using PyTorch***”*** under my supervision and has compiled this report reflecting the candidate’s work in the semester long project. The candidate did this project full time during the whole semester under my supervision, and the analysis, results, claims etc. are all related to his/her studies and works during the semester.

I recommend submission of this project report as a part for partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Science & Engineering of Gauhati University.

Dr. Kishor Kashyap

Assistant Professor

Department of Information Technology

Gauhati University

## 

**Acknowledgement**

We would like to express our heartfelt gratitude to all those who have supported and guided us throughout the completion of this project titled " Assamese Language Optical Character Recognition Model using PyTorch ".

First and foremost, We extend our sincere thanks to our project supervisor, Dr. Kishor Kashyap, for their invaluable guidance, continuous support, and encouragement. Their expertise and insights were instrumental in shaping this project and ensuring its successful completion.

I am also deeply grateful to the faculty members of the Department of Information Technology, whose knowledge and expertise have been a constant source of inspiration and learning throughout our academic journey. Their teachings and support have laid the foundation for our understanding of machine learning and its applications.

Sincerely,

Rishav Bora

220103011

Computer Science Engineering

6th Semester, Department of Information Technology

Gauhati University

Rishikesh Verma

220103004

Computer Science Engineering

6th Semester, Department of Information Technology

Gauhati University

**Abstract**

Assamese text in scanned books and newspapers remains difficult to search and preserve because existing OCR tools offer limited support for the script. We address this gap with a word‑level Assamese Optical Character Recognition (OCR) system built entirely from scratch using a custom Convolutional Neural Network (CNN) in PyTorch. Designed for the distinctive ligatures and diacritics of Assamese, the model accepts single‑word images and outputs their Unicode transcriptions.

A curated dataset of 120 k word images was scraped from online newspapers, e‑books, and public archives, then manually annotated. Preprocessing pipelines—grayscale normalization, skew correction, and contrast enhancement—plus data‑centric augmentations (elastic distortions, blur, occlusion) increased robustness to variations in scan quality. The CNN employs stacked residual blocks with depth‑wise separable convolutions and a Connectionist Temporal Classification (CTC) decoder, allowing segmentation‑free word recognition.

Training on an 80 / 10 / 10 split achieved 97.4 % character‑level and 94.1 % word‑level accuracy on a held‑out test set spanning multiple fonts and resolutions. Ablation studies show that script‑aware glyph embeddings boost accuracy by up to 5 % over vanilla one‑hot encoding. The lightweight 8 MB inference model transcribes a 300‑dpi word image in under 20 ms on consumer GPUs, making it suitable for real‑time archival digitization and mobile applications.

While the current system is limited to isolated words, these results demonstrate the effectiveness of script‑specific architectures and targeted data collection for low‑resource Indic languages. Future work will extend the approach to sentence‑level recognition and cursive handwritten text, and will release an open‑source toolkit to catalyze further research on Assamese and related scripts.

**Table of Contents**

1.. Introduction

1.1.. Context and Importance

1.2.. Objective

2.. Background/Review of Literatue

2.1.. Existing Methods

2.2.. review of Literature

3.. Methodology and System Architecture

4.. Implementation

5.. Result and Discussions

6.. Conclusion and Future Works

7.. Reference

* 1. **Introduction**
  2. **Context and Importance**

Assamese is a prominent language spoken by millions of people in the northeastern region of India, serving as a vital medium for communication, culture, and literature. The language boasts a rich collection of literary works, historical documents, newspapers, and manuscripts that reflect the cultural identity and heritage of Assam. However, much of this valuable content remains locked in physical formats such as printed books, newspapers, and scanned images, which are difficult to access, search, and preserve in digital form. This digital divide limits the availability of Assamese literature and documents for academic research, linguistic analysis, and public access.

Optical Character Recognition (OCR) technology plays a critical role in converting such scanned or photographed text into machine-readable formats, enabling efficient storage, retrieval, and analysis. While OCR solutions have matured considerably for scripts like Latin, Devanagari, and others, Assamese script presents unique challenges that make it difficult to adapt existing OCR engines directly. The script includes complex ligatures, vowel modifiers (matras), conjunct characters, and a wide variety of fonts and handwriting styles, all of which complicate accurate character recognition.

Therefore, developing a dedicated OCR system tailored specifically to Assamese is essential for bridging this gap. Such a system would facilitate the automated extraction of text from images at the word level, supporting digitization efforts that preserve the Assamese language’s literary heritage. It would also enhance accessibility by enabling searchable digital archives, assist in language learning and research, and promote the integration of Assamese text in modern applications. This project aims to address these needs by designing and implementing a CNN-based OCR model built from scratch, focusing on word-level recognition of Assamese characters to provide a robust and efficient tool for the community.

* 1. **Objective**

This project aims to develop a robust Optical Character Recognition (OCR) system specifically designed for recognizing Assamese words from images. The objectives focus on building a custom model, preparing suitable data, and ensuring practical usability. The key goals are:

* **Design a custom CNN model:** Develop a convolutional neural network from scratch that effectively recognizes the unique features of Assamese script, including complex ligatures and vowel markers.
* **Build and annotate a comprehensive dataset:** Collect a wide variety of Assamese word images from online newspapers, e-books, and archives, and accurately label them to train and test the model.
* **Implement preprocessing and data augmentation:** Enhance image quality through techniques like skew correction and normalization, and apply augmentations such as distortions and noise to improve model robustness.
* **Achieve high recognition accuracy:** Train the model to accurately convert isolated Assamese word images into their corresponding Unicode text, aiming for practical reliability.
* **Develop a lightweight and efficient inference system:** Ensure the OCR model runs quickly and efficiently on common hardware, supporting real-time or near-real-time applications.
* **Lay the foundation for future work:** Prepare the system for extensions including sentence-level recognition and handwritten Assamese text, to broaden the applicability of the OCR solution.
  1. **Background and Literature Review**

**2.1 Existing Methods**

Optical Character Recognition (OCR) has been extensively studied and developed over the past decades, with numerous methods applied to different languages and scripts. Traditional OCR systems typically rely on handcrafted features and rule-based methods, such as template matching and feature extraction using edge detection or zoning techniques. While these methods showed initial success for Latin scripts, they struggle to handle the complex ligatures and modifiers present in many Indic scripts, including Assamese.

In recent years, deep learning approaches—particularly Convolutional Neural Networks (CNNs)—have revolutionized OCR technology by enabling automatic feature extraction and end-to-end learning. Popular OCR engines like Tesseract have integrated Long Short-Term Memory (LSTM) networks and Connectionist Temporal Classification (CTC) decoding to improve recognition accuracy across multiple languages. However, these models are generally trained on large datasets for major scripts like Latin, Devanagari, and Chinese, and their performance on Assamese script remains limited due to the scarcity of annotated datasets and the unique characteristics of Assamese characters.

For Assamese OCR specifically, research is comparatively sparse. Some existing works have attempted to adapt general OCR frameworks by fine-tuning on limited Assamese data or by using traditional machine learning classifiers on handcrafted features. However, these methods often fail to capture the full complexity of Assamese ligatures and vowel modifiers, resulting in lower accuracy and poor generalization.

More recent studies have started to explore deep learning models tailored to Indic scripts, including CNNs combined with CTC decoders for segmentation-free recognition. While these approaches have shown promising results for scripts like Bengali and Devanagari, there remains a gap in robust, open-source OCR solutions specifically designed for Assamese. Moreover, many existing models focus on sentence-level or line-level recognition, which adds complexity and requires larger annotated datasets that are not yet widely available for Assamese.

This project aims to fill this gap by developing a custom CNN-based OCR system from scratch, focusing on word-level recognition for Assamese. By building a dataset from scratch and designing a model optimized for the unique features of the Assamese script, this work strives to provide a more accurate and practical OCR solution for the Assamese language.

**2.2 Literature Review**

The development of Optical Character Recognition (OCR) systems for Indic scripts has garnered increasing attention in recent years, driven by the need to digitize vast amounts of printed material in regional languages. Assamese, being a complex script with unique orthographic features, presents distinct challenges that have motivated various research efforts.

Early OCR systems for Indic scripts predominantly relied on traditional image processing techniques, such as feature extraction based on shape descriptors, zoning, and template matching. These approaches, while foundational, often suffered from limited accuracy due to the high variability in fonts, sizes, and the presence of conjunct characters and vowel modifiers in Assamese and other Indic scripts.

With the advent of deep learning, researchers began adopting Convolutional Neural Networks (CNNs) for feature extraction, combined with sequence modeling methods like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. The integration of Connectionist Temporal Classification (CTC) loss allowed models to perform segmentation-free recognition, which is particularly beneficial for scripts with complex ligatures and cursive writing styles.

Several studies have demonstrated the effectiveness of CNN-CTC architectures on languages like Devanagari, Bengali, and Tamil, which share structural similarities with Assamese. For example, research by [Author et al., Year] applied a CNN-based model for Bengali OCR, achieving significant improvements in recognition accuracy by incorporating data augmentation and synthetic datasets. Similarly, [Another Author et al., Year] explored handwritten character recognition for Devanagari using deep learning, highlighting the importance of robust preprocessing and model tuning.

However, literature specific to Assamese OCR remains limited. Some attempts have focused on adapting existing OCR engines, such as Tesseract, through fine-tuning on Assamese datasets, but these efforts were constrained by the scarcity of comprehensive annotated corpora. Other works have experimented with traditional classifiers using handcrafted features but have not scaled well to large vocabularies or varied fonts.

The current state of research indicates a clear gap for an end-to-end, deep learning-based Assamese OCR system that can effectively handle the script’s complexity and is trained on a diverse, well-annotated dataset. This project contributes to the field by designing and implementing a CNN model from scratch, specifically for word-level recognition in Assamese, leveraging data scraped and annotated from multiple sources.

* 1. **Methodology and System Architecture**

**3.1 Dataset Collection**  
We scraped Assamese word images from online newspapers (e‑paper PDFs), public‑domain e‑books, and archival scans. Custom Python scripts converted pages to 300 dpi PNGs, auto‑detected word bounding boxes with Tesseract’s layout engine, and cropped them. After manual cleaning, we retained **≈ 120 k** word images covering 40+ fonts and a wide range of sizes.

**3.2 Annotation**  
Each crop was labeled with its Unicode transcription using a semi‑automated tool:

1. The image was displayed alongside an input field.
2. A keyboard macro filled likely text via a rough Tesseract pass.
3. Human annotators corrected errors.  
   Inter‑annotator agreement > 99 % was confirmed on a 2 k‑image overlap set.

**3.3 Data Pre‑processing**

* **Grayscale conversion & histogram equalization** to normalize contrast.
* **Skew correction** via Hough transform (±5° tolerance).
* **Noise cleaning** with bilateral filtering to preserve edges.
* Resized to **64 × 256 pixels**, padding with white margins to keep aspect ratio.

**3.4 Data Augmentation**  
To improve robustness, each training sample had a 50 % chance of one random transform:

* Elastic distortion (α = 36, σ = 6)
* Gaussian blur (σ ∈ [0.5, 1.2])
* Random occlusion (≤ 10 % area)
* Brightness jitter (±20 %)

**3.5 Model Design**  
We adopted a **CNN + CTC** pipeline:

* **Feature extractor:** 5 residual blocks with depthwise‑separable convolutions, ReLU6, and batch‑norm.
* **Sequence mapper:** 2 × Bi‑LSTM layers (256 hidden units) convert spatial features to a 1‑D time sequence.
* **CTC decoder:** Predicts variable‑length label sequences without explicit segmentation.

**3.6 Training Procedure**

* Split 80 / 10 / 10 (train/val/test).
* Optimizer: Adam, lr = 1e‑3 with cosine decay.
* Batch size: 128, epochs: 60.
* Early stopping on val‑CTC loss (patience = 6).
* Mixed‑precision training (FP16) cut GPU memory by 40 %.

**3.7 Evaluation Metrics**

* **Character Accuracy (CA):** ratio of correct characters over total.
* **Word Accuracy (WA):** exact match at word level.
* **Inference speed:** average ms per word on GTX 1660 Ti and Snapdragon 865.