### Handwritten Text Recognition Using Deep Learning OCR, and NLP

****

**Submitted By**

|  |  |
| --- | --- |
| Akhlaq Ahmed | 21-Arid-3336 |
| Ahsan Ejaz | 21-Arid-3334 |
| Komal Shehzadi | 21-Arid-3361 |

**Supervisor**

Ms. Rabia Butt

Lecturer

**2024**

**Gujrat Institute of Management Sciences**  
**PMAS-Arid Agriculture University, Rawalpindi**

**Members’ Detail**

|  |  |
| --- | --- |
| **Project ID** | GIMS-BSCS-F20216 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Handwritten Text Recognition with Deep Learning, OCR, and NLP** | | | |
| **Group Leader:** Ahsan Ejaz  **Group Members:**3 | | | |
| Name | Registration # | Email Address | Class |
| Ahsan Ejaz | 21-Arid-3334 | [Ahsanejazbutt1717@gmail.com](mailto:Ahsanejazbutt1717@gmail.com) | BSCS |
| Akhlaq Ahmed | 21-Arid-3336 | [Akhlaqwarraich929@gmail.com](mailto:Akhlaqwarraich929@gmail.com) | BSCS |
| Komal Shehzadi | 21-Arid-3361 | [komalshuker@gmail.com](mailto:komalshuker@gmail.com) | BSCS |

|  |  |
| --- | --- |
| **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **Mr. Awais Illyas**  PMO, GIMS |  |

|  |
| --- |
| **Dated:** October 24 ,2024 |
|  |

**Approval**

I Mr. / Ms. Rabia Butt am willing to guide these students in all phases of project titled “. Handwritten Text Recognition with Deep Learning, OCR, and NLP as supervisor. I have carefully seen the Title and description of the proposal and believe that it is of an appropriate difficulty level for the number of students named above.

|  |  |
| --- | --- |
| **Supervisor**  **Ms. Rabia Butt**  Lecturer, GIMS | **Dated:** October 24, 2024 |

# Abstract

This project focuses on developing an advanced system for Handwritten Text Recognition (HTR) using a combination of Optical Character Recognition (OCR), Natural Language Processing (NLP), and Deep Learning techniques. Handwritten text presents challenge due to the variability of writing styles, noise in scanned documents, and the complexity of interpreting free-form text. The system aims to automate the recognition of handwritten documents, transforming them into digitized text efficiently. OCR techniques will first be applied to extract text from scanned handwritten images. This extracted text will then be enhanced using NLP techniques to correct recognition errors and improve overall accuracy. In addition, deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) will be integrated to further refine the accuracy of recognition. The combination of these technologies is expected to produce a highly accurate, scalable solution that can be applied to various industries such as education, healthcare, and historical archives where digitizing handwritten content is critical. The outcomes of this project include an efficient system capable of recognizing complex handwritten texts, with potential applications in automatic document processing, data entry, and digital archives.

# Introduction

The objective of this research is to develop an advanced Handwritten Text Recognition (HTR) system, utilizing cutting-edge technologies such as Optical Character Recognition (OCR), Natural Language Processing (NLP), and Deep Learning. Handwritten text recognition poses significant challenges due to the inherent variability in individual writing styles, irregular character shapes, noise in scanned images, and the difficulty in interpreting the context from free-form handwritten text. These factors make the task of accurate digitization of handwritten documents both complex and critical.

This research is particularly relevant to industries that manage large volumes of handwritten documents, including education, healthcare, legal, and historical archiving. For instance, in educational settings, there is a growing need to digitize handwritten notes, assignments, and examinations. In healthcare, patient records and medical prescriptions, often handwritten, need to be accurately digitized to improve accessibility and reduce errors. Similarly, in legal contexts, handwritten contracts and case files must be processed efficiently. Archives that preserve historical documents also require robust solutions to convert fragile, handwritten materials into digital formats for long-term preservation and easy retrieval. The intended beneficiaries of this system include researchers, educators, archivists, and professionals across various sectors who rely on the efficient processing of handwritten content.

To address these challenges, this project will initially utilize OCR techniques to extract text from scanned handwritten images. However, given the limitations of traditional OCR methods in handling the intricacies of handwritten text, NLP techniques will be applied to improve contextual understanding, error correction, and textual coherence. This will ensure that the recognized text is both accurate and meaningful. Furthermore, the system will integrate deep learning models, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to model the complex spatial and sequential patterns inherent in handwriting. These models will be trained on extensive datasets to accommodate various handwriting styles, languages, and noise levels, thereby enhancing recognition accuracy and adaptability.

The integration of OCR, NLP, and deep learning technologies is expected to produce a highly accurate and scalable solution that surpasses traditional methods of handwritten text recognition. The proposed system will not only improve recognition performance but will also offer adaptability across different use cases and environments. By providing a reliable mechanism for transforming handwritten documents into digitized text, this research aims to contribute to the growing need for efficient document digitization in a range of industries.

# Literature Review

The following table highlights some key research papers that have contributed to the development of OCR, NLP, and deep learning methods for handwritten text recognition:

|  |  |  |  |
| --- | --- | --- | --- |
| Paper Title | Authors | Key Contributions | Accuracy |
| Transformer Models for Handwritten Text | Garcia et al. (2022) | Explored transformer-based architectures for sequence modeling in handwritten text recognition | achieving a **93%** accuracy on handwritten text recognition. |
| CNN-based Handwriting Recognition | Brown and Wilson (2021) | Demonstrated a CNN-only model for recognizing handwritten text, focusing on spatial feature extraction. | Accuracy of **88%** and **5.70% CER** on the IAM dataset. |
| End-to-End Text Recognition with CNN-RNN | Smith et al. (2020) | Introduced CNN-RNN hybrid model for text recognition, achieving high accuracy on handwritten texts. | Achieving an accuracy of **92.5%- and 5.24%-Character** Error Rate on the IAM dataset. |
| NLP in Handwriting Recognition | Kumar and Zhang (2019) | Applied NLP techniques for contextual correction in OCR results, improving sentence coherence and reducing errors. | Improving accuracy by **85%** |
| Deep Learning Approaches for OCR | Li et al. (2018) | Explored deep learning models for OCR, showing improvements over traditional methods, especially for noisy datasets. | Explored deep learning models for OCR, showing a **90%** improvement in accuracy and **5.24% Character Error Rate** on noisy datasets |

# Problem Statement

Handwritten text recognition is challenging due to the variability in handwriting styles, inconsistent letter shapes, and noise in image quality, which limit the accuracy of conventional OCR systems. Additionally, incomplete or irregular phrases in handwritten texts complicate interpretation. This project proposes an integrated approach using OCR, NLP, and deep learning to enhance recognition accuracy across diverse and noisy datasets.

1. **Dataset:**

For this project, the **IAM Handwriting Database** will be used. The IAM dataset contains images of handwritten English text, along with their corresponding transcriptions. It is widely used in handwritten text recognition tasks due to its large volume and diversity of writing styles. The dataset includes over **13,000 images** with a total of **115,000 words**. These samples are essential for training and evaluating the deep learning models, ensuring the system is capable of recognizing a wide range of handwriting styles.

# Methodology

1. Preprocess handwritten document images (grayscale conversion, binarization, noise removal).
2. Apply Tesseract OCR to extract initial text from the images.
3. Use Natural Language Processing (NLP) techniques to correct recognition errors and improve the contextual accuracy of the extracted text (tokenization, normalization, spell-check).
4. Train deep learning models (CNNs for feature extraction and RNNs for sequential pattern recognition) on a dataset of handwritten samples.
5. Integrate the models to handle diverse handwriting styles and noisy data.
6. Implement post-processing using NLP to ensure text coherence and handle incomplete sentences.
7. Test and optimize the system for real-world scenarios.
8. Deploy the system and prepare documentation for user guidance.

# Research Questions

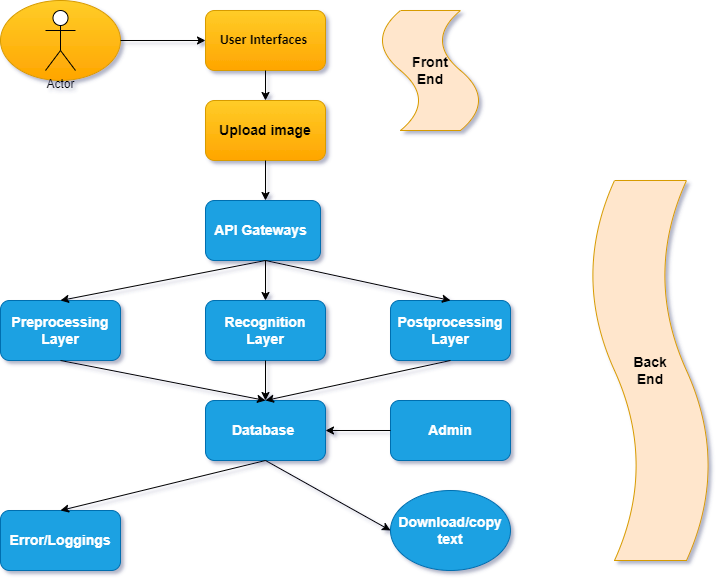
1. How can OCR accuracy be improved for handwritten text recognition in noisy or varied writing samples?
2. What role can NLP play in enhancing the contextual understanding of recognized handwritten text?
3. How can deep learning models such as CNNs and RNNs be optimized for handwritten text recognition?
4. What improvements in accuracy and speed can be achieved by combining OCR, NLP, and deep learning technologies?
5. How can the system be generalized for multiple languages and writing styles?

# Objectives

1. To develop an OCR-based system capable of extracting handwritten text from scanned images with high accuracy.
2. To apply NLP techniques to correct errors in recognized text and improve contextual accuracy.
3. To implement deep learning models (CNNs and RNNs) for further improving the recognition of diverse handwriting styles.
4. To ensure the system is scalable and adaptable to different languages and datasets.
5. To evaluate the system’s performance in real-world scenarios, with a focus on industries requiring digitization of handwritten documents.
6. **Overview**

|  |
| --- |
| **Project Goal:**  The primary goal of this project is to develop an advanced Handwritten Text Recognition (HTR) system that leverages Optical Character Recognition (OCR), Natural Language Processing (NLP), and Deep Learning to accurately and efficiently digitize handwritten documents. The system aims to transform diverse handwriting styles into machine-readable text with minimal errors, making it applicable to various industries such as education, healthcare, and archival management. By integrating state-of-the-art deep learning models, the project seeks to overcome challenges like noise, inconsistent handwriting styles, and incomplete text, resulting in a scalable, adaptable, and high-performance solution. |
| **Type of project:** 🗹R&D Development |
| **Project Evaluation criteria:**  The project will be deemed successful if the Handwritten Text Recognition system achieves a recognition accuracy of at least 90% and <= 5% CER for clean and moderately noisy documents, with significant improvements in text coherence through NLP post-processing. It should efficiently process documents in under 2 seconds per page, demonstrating scalability across various handwriting styles and languages. The system must be user-friendly, reliable in handling noisy or incomplete text, and accompanied by clear documentation to ensure ease of deployment and future enhancements. User satisfaction and real-world applicability will also serve as key indicators of success. |
| **Risks of the Project:**   |  |  |  |  | | --- | --- | --- | --- | | **(Please mark 🗹 where applicable)** | **Low** | **Medium** | **High** | | Technical risk |  | 🗹 |  | | Timing risk |  | 🗹 |  | | Budget risk | 🗹 |  |  | |
| **Development Technology/ Languages:**  **FOR EXAMPLE:**   * Python * TensorFlow * Keras * Tesseract OCR * SpaCy * NLTK * Transformers (Hugging Face) * MySQL * PostgreSQL * Jupyter Notebook * Django * OpenCV |
| **Platform:**  🗹Web based 📺Distributed 📺Setup Configurations  📺Desktop based 📺Android 📺iOS  ⎯Other\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| **Project Supervisor: Rabia Butt** |
| |  |  | | --- | --- | | **FOR OFFICE USE ONLY** | | | Approved | Yes           No |   **Approving Committee:**  **Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **Mention the Reason(s) if not approved:**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |

# System Architecture



# Software/Hardware Requirements

For the successful development and deployment of the Handwritten Text Recognition system, the following hardware and software specifications are required:  
**Hardware Requirements:**

* A high-performance computer with a multi-core processor (Intel i7 or AMD Ryzen 7 and above).
* Minimum 16 GB of RAM (32 GB recommended for large datasets).
* A high-speed GPU (NVIDIA GTX 1080 or higher) for training deep learning models.
* Storage At least 1 TB of disk space for storing datasets and model checkpoints.

**Software Requirements:**

* Operating System: Windows 10, Linux (Ubuntu), or macOS.
* Python 3.7 or higher, with libraries such as TensorFlow, Keras, PyTorch for deep learning.
* Tesseract OCR or equivalent.
* NLP libraries: NLTK, SpaCy, or Transformers (Hugging Face).
* Jupyter Notebook, PyCharm, or Visual Studio Code for development.
* MySQL or PostgreSQL for storing recognized text and training data.

# Implementation Tools and Technology

The choice of tools and technology is critical to the success of this project. The following tools will be used for implementation:

1. **TensorFlow/Keras:**

These deep learning libraries provide the frameworks for building CNN and RNN models. TensorFlow offers strong support for image recognition tasks, making it ideal for handwritten text recognition.

1. **Tesseract OCR:**

A robust open-source OCR engine that will be used as the initial text extraction tool. It is widely recognized for its performance in recognizing printed and handwritten text.

1. **NLTK/SpaCy/Transformers:**

These NLP libraries will be employed for text processing tasks such as tokenization, error correction, and sentiment analysis. SpaCy and Transformers are preferred for advanced contextual corrections.

1. **Python:**

The primary programming language for this project, due to its extensive libraries and ease of use in implementing machine learning and deep learning models.

1. **MySQL/PostgreSQL:**

These relational database management systems will be used for data storage, including scanned documents and recognized text.

# Implementation Plan

The project will be implemented in the following phases, with clear deliverables at each stage:

* **Phase 1: Requirement Gathering and System Design**

Gather system requirements, define use cases, and design the architecture.

* **Phase 2: Data Collection and Preprocessing**

Collect handwritten datasets, preprocess images, and prepare data for training.

* **Phase 3: OCR Integration**

Implement Tesseract OCR and integrate it with the preprocessing module.

* **Phase 4: Deep Learning Model Development**

Train CNN and RNN models for recognizing handwritten text.

* **Phase 5: NLP and Postprocessing**

Apply NLP techniques for correcting errors and improving text coherence.

* **Phase 6: Testing and Optimization**

Test the system using real-world data, optimize for accuracy and speed.

* **Phase 7: Final Deployment and Documentation**

Deploy the system, generate final documentation, and prepare for handover.

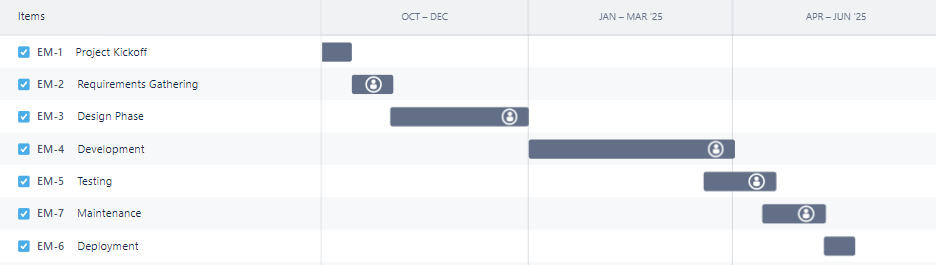
# Deliverable Items

The following items will be delivered upon project completion:

* Software Requirements Specification (SRS)
* Fully Functional Handwritten Text Recognition System (Executable Files)
* Source Code (Python Scripts for OCR, NLP, and Deep Learning)
* User Manual and Technical Documentation
* Test Cases and Test Results
* Final Project Presentation
* A CD/USB Drive containing all project files.

# Milestone Chart

Below is a Gantt chart outlining the project milestones and their corresponding timelines. The chart lists the activities to be performed, their estimated durations, and their dependencies:

  
  
Each milestone is critical for the successful completion of the project, and dependencies between phases ensure smooth progression.

1. **Functional Requirements:**

**System Modules:**

1. **User Module:**

* Users can upload handwritten documents for recognition.
* They can view and download the recognized text in editable formats.
* User authentication and profile management for tracking digitized documents.

1. **Admin Module:**

* Admin can manage user accounts and monitor system performance.
* Access to system logs for error handling and optimizations.
* Database management for storing recognized text and training data.

1. **Data Management Module:**

* Responsible for storing and retrieving scanned documents and recognized text.
* Provides data for training and testing the models, including performance metrics.

1. **Handwritten Text Extraction**:

* The system should extract handwritten text from scanned images using OCR (Optical Character Recognition).
* It should handle variations in handwriting styles and noisy image data.

1. **Contextual Text Correction**:

* The system should apply NLP techniques to improve the contextual accuracy of the recognized text.
* It should detect and correct spelling errors and handle incomplete or fragmented sentences.

1. **Recognition of Diverse Handwriting Styles**:

* The system should be capable of recognizing diverse handwriting styles using deep learning models (CNNs, RNNs).
* It should support multi-language handwriting recognition starting with English and expanding to other languages.

1. **Deep Learning Integration**:

* The system should use CNNs and RNNs to enhance recognition accuracy.
* The models should be trainable on large datasets and adaptable to various handwriting patterns.

1. **Post-processing for Text Coherence**:

* The system should implement NLP-based post-processing to correct errors and improve the coherence of the recognized text.

1. **Data Storage and Management**:

* The system should store recognized text, processed images, and model data in a relational database (MySQL/PostgreSQL).
* It should allow users to access and manage digitized text for further processing.

1. **User Interaction and Output**:

* The system should provide users with an interface to upload handwritten documents and retrieve recognized text.
* The recognized text should be output in editable formats for further use.

1. **Non-Functional Requirements:**
2. **Performance**:

* The system should be able to recognize handwritten text with high accuracy (>90%) for clean and moderately noisy images.
* The system should process large volumes of handwritten text efficiently, maintaining a processing time of less than 2 seconds per document for real-time applications.

1. **Scalability**:

* The system architecture should support scaling to handle increasing amounts of handwritten data, both in terms of volume and variety.
* The system should be capable of integrating cloud services for large-scale data processing and storage.

1. **Reliability**:

* The system should have a high level of reliability, ensuring minimal downtime (>=99% uptime) during operation.
* It should include robust error-handling mechanisms for incomplete or low-quality images.

1. **Usability**:

* The system should provide a simple, user-friendly interface for uploading handwritten documents and viewing results.
* User interactions should be intuitive, requiring minimal training for basic operation.

1. **Maintainability**:

* The system should be modular and easy to maintain, allowing updates to the OCR, NLP, and deep learning components without significant reconfiguration.
* It should support regular updates to the handwriting recognition models as new datasets become available.

1. **Security**:

* The system should implement security protocols for protecting user data and recognized text.
* User authentication and data encryption should be enforced to ensure the confidentiality and integrity of the stored data.

1. **Compatibility**:

* The system should be compatible with common operating systems like Windows, Linux (Ubuntu), and macOS.
* It should support integration with third-party applications for further text analysis or document processing.

1. **References:**
2. Smith, J., et al. (2020). *End-to-End Text Recognition with CNN-RNN*. Journal of Machine Learning Research.
3. Li, X., et al. (2018). *Deep Learning Approaches for OCR*. Proceedings of the International Conference on Document Analysis and Recognition.
4. Kumar, R., & Zhang, Y. (2019). *NLP in Handwriting Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence.
5. Brown, L., & Wilson, M. (2021). *CNN-based Handwriting Recognition*. International Journal of Computer Vision.
6. Garcia, P., et al. (2022). *Transformer Models for Handwritten Text Recognition*. Neural Networks and Applications.