**White Report: Enhancing OCR Accuracy on Bank Forms Using GEN AI**

**Introduction**

This document consolidates our extensive work on improving Optical Character Recognition (OCR) accuracy for bank form images. The focus of our project was to preprocess bank form images by removing interfering graphical elements and standardizing the text color and background. Bank forms, designed with various visual aids like lines, grids, and checkboxes, pose significant challenges for OCR systems, leading to incorrect text extraction and reduced accuracy. This report details the tasks undertaken, challenges faced, and progress made in developing a robust preprocessing pipeline and training models to achieve higher OCR accuracy and efficiency.

## Overview of Tasks and Approaches

## Common Objectives

## Remove Graphical Elements: The primary task involves eliminating dotted, vertical, and horizontal lines, as well as checkboxes from the forms. These elements, while useful for human users, can confuse OCR systems and reduce their accuracy.

## Standardize Text and Background: To enhance the clarity and contrast of the text for OCR, we aim to convert all text to black and change the background color to white. This standardization helps in creating a uniform input for OCR systems, improving their ability to accurately recognize and extract text.

## Enhance OCR Accuracy: By removing unnecessary elements and standardizing the text and background, we strive to improve the clarity of the text in the images. This enhancement is crucial for achieving higher OCR accuracy and reliability in text extraction processes.

## Techniques and Tools

## OpenCV: We utilize OpenCV for various image preprocessing tasks, including the removal of lines and checkboxes. OpenCV offers a range of functions that help in manipulating and enhancing images to meet our specific requirements.

## OCR Models: To handle text extraction, we experiment with several OCR models, including TesseractOCR, EasyOCR, and PaddleOCR. Each model has its strengths, and we assess their performance in extracting text from preprocessed images.

## Generative AI Models: CycleGAN and Pix2Pix are employed to further enhance the quality of images post-preprocessing. These models help in refining the images by learning the patterns of graphical elements and improving the overall text clarity.

## Image Superimposition: This technique involves superimposing checkboxes and dotted lines on clean images to create a training dataset. By using these paired images (superimposed input and original ground truth), we train our models to recognize and remove these elements effectively.

## Hardware and Software Requirements

* GPU-enabled system for faster OCR processing.
* High-resolution scanner for capturing document images.
* Python 3.x
* Libraries: OpenCV, EasyOCR, TesseractOCR, PaddleOCR, NumPy, Matplotlib, File handling libraries (os).

**Detailed Weekly Summaries**

**Week 1: Initial Exploration and Dataset Collection**

**Tasks in Bucket**

In the first week, our primary focus was on understanding the different OCR models and preprocessing techniques available. We started by collecting a diverse dataset of bank forms to work with. This dataset included various forms with different designs and graphical elements to ensure a comprehensive training and testing process.

**Planned Tasks:**

Our planned tasks involved collecting and annotating the dataset to make it suitable for training and testing. We also aimed to perform initial tests using PyTesseract and OpenCV without cropping the images. The goal was to assess the baseline performance of these tools in removing graphical elements and extracting text.

**End of Week Status:**

By the end of the week, we successfully collected and annotated the initial dataset. However, our experiments with PyTesseract and OpenCV revealed that PyTesseract struggled to remove graphical elements effectively, and OpenCV, while capable of removing some lines, resulted in unclear text.

**Summary of Work:**

During this week, we explored various OCR models and preprocessing techniques to understand their capabilities and limitations. We faced challenges in removing graphical elements without affecting text clarity, highlighting the need for more advanced preprocessing methods.

**Week 2: Using OpenCV for Preprocessing**

**Tasks:**

In the second week, we focused on improving our preprocessing techniques using OpenCV. The goal was to enhance the removal of lines, grids, and checkboxes while preserving the text clarity in the images.

**Planned Tasks:**

We planned to implement various OpenCV techniques, such as morphological operations and contour detection, to remove the graphical elements more effectively. Additionally, we aimed to test these techniques on our entire dataset to evaluate their performance comprehensively.

**End of Week Status:**

By the end of the week, we observed improvements in line removal using OpenCV. However, maintaining text clarity remained a challenge, as some preprocessing operations still caused the text to blur.

**Summary of Work:**

Throughout this week, we applied different OpenCV techniques to enhance image preprocessing. While we achieved better results in removing graphical elements, preserving text clarity was still problematic, indicating the need for further refinement and additional methods.

**Week 3: Experimenting with Generative AI Models**

**Tasks:**

In the third week, we started experimenting with generative AI models, specifically CycleGAN and Pix2Pix, to enhance the images after initial preprocessing with OpenCV. These models have the potential to learn and improve the image quality by removing residual graphical elements and refining text clarity.

**Planned Tasks:**

We planned to train CycleGAN and Pix2Pix on our preprocessed dataset. The training involved using paired images (input with graphical elements and clean ground truth) to teach the models to remove unwanted elements and enhance text clarity.

**End of Week Status:**

By the end of the week, we found that CycleGAN did not provide optimal results, while Pix2Pix performed better, especially when the images were cropped to 256x256 pixels. However, some lines and checkboxes remained, indicating the need for further tuning.

**Summary of Work:**

This week, we focused on training CycleGAN and Pix2Pix on preprocessed images. We realized the importance of image resolution in training generative models and the need for further adjustments to achieve complete removal of graphical elements and clear text.

**Week 4: Combining OCR and Preprocessing Techniques**

**Tasks:**

In the fourth week, we combined OCR and preprocessing techniques to achieve better results. We integrated TesseractOCR and EasyOCR with OpenCV to improve text retention while removing graphical elements.

**Planned Tasks:**

We planned to crop images to 256x256 pixels and preprocess them using OCR models and OpenCV. Additionally, we employed the superimposition technique to enhance the training dataset by creating images with superimposed graphical elements and their corresponding clean ground truth.

**End of Week Status:**

By the end of the week, combining these techniques improved the results, but some text was still missed. The superimposition technique showed promise in enhancing the training dataset, but further refinement was needed.

**Summary of Work:**

This week, we integrated OCR models with OpenCV for preprocessing. The superimposition technique helped in training the models to remove graphical elements more effectively, but we faced challenges in achieving complete text retention.

**Week 5: Implementing PaddleOCR**

**Tasks:**

In the fifth week, we introduced PaddleOCR for text detection and preprocessing. PaddleOCR is a state-of-the-art OCR model known for its accuracy and efficiency in text extraction.

**Planned Tasks:**

We planned to preprocess images with PaddleOCR by detecting text and placing it on a white background. The preprocessed images were then paired with the ground truth to create a robust training dataset.

**End of Week Status:**

By the end of the week, PaddleOCR improved text detection, but the preprocessing still required refinement to ensure complete removal of graphical elements.

**Summary of Work:**

This week, we utilized PaddleOCR for text detection and preprocessing. The results showed improved text clarity, but further work was needed to refine the preprocessing techniques to ensure all graphical elements were removed effectively.

**Week 6: Refining and Validating Models**

**Tasks:**

In the final week, we focused on fine-tuning our preprocessing techniques and validating the models on new datasets. The goal was to ensure that our methods were robust and generalizable across different forms.

**Planned Tasks:**

We planned to implement advanced preprocessing techniques to enhance data quality. Additionally, we aimed to fine-tune the model parameters and validate the models on diverse datasets to ensure they performed well across different scenarios.

**End of Week Status:**

By the end of the week, advanced preprocessing techniques significantly improved data quality, and the models demonstrated good performance on diverse datasets, indicating their robustness and generalizability.

**Summary of Work:**

This week, we focused on refining preprocessing techniques and validating our models. The improvements in data quality and model performance showed the effectiveness of our methods, although continuous refinement is necessary to handle all possible variations in bank forms.

**Challenges Faced**

 **Understanding Training Methods**: It was challenging to grasp the various training methods and their applicability across different models. Understanding the strengths and limitations of each method required extensive experimentation and analysis.

 **Maintaining Accuracy**: Achieving and maintaining high accuracy in predictions was difficult, especially when dealing with diverse input images. Ensuring that the preprocessing did not compromise text clarity while removing graphical elements required careful balancing.

 **Preprocessing Issues**: Removing unwanted elements from images consistently was challenging. Some preprocessing techniques blurred the text, making it difficult for OCR models to extract information accurately. Continuous tweaking and refinement were necessary to address these issues.

**Progress and Achievements**

 **Graphical Element Removal**: We successfully removed most graphical elements from the images, such as lines and checkboxes. This was achieved through a combination of OpenCV techniques and generative AI models.

 **Improved OCR Accuracy**: By standardizing text color and background, we significantly improved OCR accuracy. The clarity of the text in the preprocessed images facilitated better text extraction by OCR models.

 **Robust Preprocessing Pipeline**: We developed a robust preprocessing pipeline that integrates multiple techniques to ensure the removal of unwanted elements and retention of clear text. This pipeline is flexible and can be adapted to different types of forms.

**Conclusion**

The six-week project demonstrated significant progress in preprocessing bank form images for enhanced OCR accuracy. Despite challenges, the combination of advanced preprocessing techniques and generative AI models led to substantial improvements. Our efforts to understand and refine various methods resulted in a robust and effective preprocessing pipeline. Moving forward, continuous refinement and validation are necessary to handle the diversity of bank forms and further improve OCR accuracy. This project highlights the potential of integrating multiple techniques to overcome complex challenges in image preprocessing and text extraction.