UNIVERSITY OF MAKATI

Development of VisionAid: Real-Time Document and Natural Scene Text Recognition with Text-to-Speech Assistance for Visually Impaired Individuals using YOLO v9

A THESIS SUBMITTED TO THE FACULTY OF COLLEGE OF COMPUTING

AND INFORMATION SCIENCES IN CANDIDACY FOR THE DEGREE OF

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

(COMPUTATIONAL AND DATA SCIENCES ELECTIVE TRACK)

DEPARTMENT OF COMPUTER SCIENCE

BY

CAPARAS, ALLEN JAY A.

DE GUZMAN, CHEN ABRIEL D.

QUESTERIA, ALVEN C.

MAKATI CITY, PHILIPPINES

APRIL 2024

The THESIS entitled:

Development of VisionAid: Real-Time Document and Natural Scene Text Recognition with Text-to-Speech Assistance for Visually Impaired Individuals using YOLO v9

submitted by Allen Jay A. Caparas, Chen Abriel D. De Guzman, and Alven C. Questeria has been examined and is recommended for Oral Defense.

ASST.PROF. ALI A. NAIM, MIS
Chair, IS and ITE Allied Programs

ROMMEL T. GARMA, Ph.D
Adviser

JOEL B. MANGABA, DT.
Dean
College of Computing and Information Sciences

The Faculty of the Department of Computer Science, College of Computing and Information Sciences, University of Makati, ACCEPTS THE THESIS entitled:

Development of VisionAid: Real-Time Document and Natural Scene Text Recognition with Text-to-Speech Assistance for Visually Impaired Individuals using YOLO v9

submitted by Allen Jay A. Caparas, Chen Abriel D. De Guzman, and Alven C. Questeria, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science (Computational and Data Sciences Track).

ASST.PROF. CHRISTIAN MICHAEL M. MANSUETO, MIS-ITEM
Panel Member

MR. JADE IRVINE MAGPAYO
Panel Member

ASST.PROF. ALI A. NAIM, MIS
Panel Chair

ABELARDO T. BUCAD, MIS-ITEM
Technical Adviser

JOEL B. MANGABA, DT.
Dean
College of Computing and Information Sciences

Chapter 1

THE PROBLEM AND ITS BACKGROUND

Introduction

Inspired by the profound difficulties faced by visually impaired individuals in accessing textual information, the group of students known as Oculi (Latin for “eye”) began developing the app VisionAID Real-Time Image-to-Speech Assistance for Visually Impaired Individuals using Computer Vision. Driven by a shared ambition to empower those with visual impairments through innovative, accessible technology, Oculi’s VisionAid seeks to address the accessibility challenges in the Philippines. Many people in the Philippines have bad eyesight, according to The 2019 Philippine National Blindness Survey by the Department of Health (DOH) reveals that about 1.98% of the population in the Philippines.Nonetheless, the increase in the use of mobile phones, which is now considered a modern basic need for the lowest household income bracket, has redefined accessibility. VisionAid is working to address this very challenge by utilizing the availability of mobile devices to offer a practical and cost-effective solution for the visually impaired individuals.

According to Lancet Global Health Commision highlights that these prescription glasses lack affordability and are not often considered a luxury for those who live in low income countries like the Philippines(Burton, 2020). Existing solutions are costly, intricate, or unsuitable for communities of users as a result of their socio-economic backgrounds. Additionally, there is a current absence of mobile applications that utilize real-time, mobile-based image recognition, text extraction and speech synthesis that is also effective and user-friendly. This emphasizes the fact that there is a need to develop a solution that can be implemented on ordinary and commonly used mobile devices with the capability to offer accurate and reliable support in real situations.

VisionAid's goal is to provide a real time text-to-speech that can assist not only the visually impaired individual but also it creates an innovative and accessible environment. With these improvements in accessibility can increase independence and confidence among the users. VisionAid seeks to help the visually impaired individuals read text documents and natural scenes to improve their quality of life (QoL).

Background of the Study

Vision is one of the important senses we human have as it is the primary way to receive information in the environment. It allows us to see colors, shapes, distance and movements as it is crucial for navigating through physical spaces. Visual input is one of the ways for learning, especially for children to help develop their motor skills and it is effective for communication that can recognize facial expressions, body language and other non-verbal signals.

In order to be classified as normal vision, you should have a 20/20 vision, meaning you should be able to see an object clearly from 20 feet away (Cleveland Clinic, 2022). There are multiple types of visual impairment such as cataracts, uncorrected fractional error (nearsighted, and farsighted), glaucoma, and maculopathy. According to the 2019 Philippine National Blindness Survey (DOH), 1.98% of the population in the Philippines has vision impairment or blindness. This equates to 1.1 million Filipinos having cataracts, 400,000 with an uncorrected fractional error, 300,000 with glaucoma, and 200,000 with maculopathy (Villanueva, 2022).In order to be classified as Persons with Disabilities (PWD) for vision in the philippines you need to have the following, 20/70 vision, field of vision is less than 20 degrees wide in the better eye, vision cannot be improved by eye glasses, medication or surgery (ROQUE Eye Clinic, 2024).

According to the Philippine Statistics Authority, the poverty incidence among the population in the first semester of 2023, or the proportion of poor Filipinos whose per capita income is insufficient to cover their basic food and non-food needs, was estimated at 22.4 percent, or 25.24 million Filipinos. Subsistence incidence among Filipinos, or the proportion of Filipinos whose income is insufficient to cover even basic food needs, was recorded at 8.7 percent, or approximately 9.79 million Filipinos(Philippine Statistics Authority, 2023).

With mobile phones being classified as one of our modern necessities, it’s evident that they play a crucial role in connecting people nowadays. According to the data on mobile user operating systems in the Philippines as of January 2024. 85.19% of all users in the Philippines use Android, whereas 14.31% use iOS (Statista, 2024). This means that Android is used by more people in the Philippines than iPhones.

Given the significant impact of vision impairment on individuals' lives and the financial challenges faced by many in the country, utilizing artificial intelligence (AI) presents a possible solution to this problem. By having a tool designated to cater visually impaired individuals, allowing them to see things that normally cannot. We intend to find a solution in this existing gap, and improve their quality of life.

Objective of the Study

The objective of this study is to develop a real-time document and scene text recognition system with text-to-speech assistance, specifically designed for visually impaired individuals. This study needs to understand the challenges of the visually impaired individuals and provide them with auditory feedback from text documents and natural scenes. This application will have a voice command feature that can help these individuals navigate through the app. It also supports languages like Tagalog, English and Baybayin that are suitable for the people in the Philippines.

Specific Objectives

The research project aimed to achieve the following objectives:

1. **Understand Challenges**
   a. Identify the challenges faced by individuals with visual impairments, particularly in accessing written documents.
   b. Explore methods to improve their quality of life through accessible technology.
2. **Evaluate Advanced Techniques**
   a. Analyze methods in Optical Character Recognition (OCR) and machine learning to improve VisionAid's accuracy and performance.
3. **Data Preparation and Analysis**
   a. **Data Gathering, Preprocessing, and Postprocessing**: Collect, clean, and preprocess data for YOLOv9 model training.

    b. **Exploratory Data Analysis (EDA)**: Assess data patterns to ensure dataset quality for improved training outcomes.

4. **Model Development**
    a. Design and fine-tune YOLOv9 to perform layout analysis, segmentation, and OCR tasks.
    b. Develop the pipeline in Python using Google Colab, focusing on precision and time efficiency.
5. **Model Evaluation**
    a. Use key performance metrics (Mean Precision Accuracy, Precision, Recall, F1 Score) to test and evaluate the VisionAid model.
6. **Mobile Application Development**
    a. Build a mobile application integrating the trained machine learning model for real-time document and natural scene recognition via the device's camera.
7. **System Testing and Evaluation**
    a. Validate the application against **functional and non-functional requirements**.
    b. Assess the system using **ISO 25010** standards, focusing on:
        i. Functional Suitability
        ii. Performance Efficiency
        iii. Usability
        iv. Reliability
        v. Security
        vi. Compatibility
        vii. Maintainability

viii. Portability

8. **User Satisfaction Evaluation**
    a. Use a **Likert scale** to measure user satisfaction.
    b. Apply the **Technology Acceptance Model (TAM)** to evaluate user acceptance and behavior.

Scope and Limitations

This study focuses on the visually impaired individuals in the University of Makati. This VisionAid will support languages in the Philippines including Tagalog, English and Baybayin, creating a practical and user-friendly application that enhances independence and accessibility. VisionAid is limited to Tagalog, English and Baybayin, it may also not recognize voices that have a different accent than the accent model is trained on. This application will require the internet for real-time processing of text recognition and for its environmental limitations such as lighting condition, text quality and complexity.

Significance of the Study

The Development of the VisionAid, a real-time document and scene text recognition system with text-to-speech assistance, has the potential to significantly improve the lives of those who are visually impaired individuals. VisualAid aims to provide a practical, user-friendly solution that promotes independence. The significance of this study can be evaluated from the viewpoint of individuals, organization and other societal impacts.

Organization

This study provides an advanced solution that can be incorporated into assistive companies. These organizations can improve their offerings and increase their competitiveness in the market by leveraging advanced OCR , TTS and other algorithm use in the study. It provides valuable insight and development and implementation of technology focusing for visually impaired individuals, serving as a benchmark for the future innovation in the field.

Individual

VisionAid provides visually impaired individuals with the tools they need to read and comprehend text on their own from documents and natural scenes. This significantly enhances their ability to perform daily tasks, pursue education and engage in social activities. The customization of the VisualAid allows users to adjust to their preferences, providing a personalized experience that improves usability and satisfaction.

Others

Policy makers can use the study's findings to inform the policies and regulations that promote and advocate a campaign for awareness and resources dedicated to the needs of visually impaired individuals. This can lead to a more supportive environment and improve access to necessary technologies.

Raising awareness about the challenges faced by the visually impaired individuals and the technology solution fostering a more inclusive society. By prioritizing the inclusion of the visually impaired individuals, we not only acknowledge the hardship the barriers they faced but also emphasize the developing accessible technology.The study highlights the importance of accessibility in technology development, promoting a culture of inclusion and innovation that benefits everyone.

The researchers can enhance their academic's knowledge and professional growth and can also get an opportunity into technology by developing a VisionAid application. This study empowers the research to be open for inclusivity of the visually impaired individuals to create independence and a more innovative society.

Operational Definition of Terms

Baybayin – An ancient Filipino script used before the Spanish colonization. In the paper, Baybayin is supported as a language, allowing visually impaired individuals to access the ancient script.

Cataracts – A medical condition where there is a cloudy area in the lens, leading to blurred vision. VisionAid aims to assist these individuals by providing auditory output from text and natural scenes.

Convolutional Neural Network (CNN) – A type of deep learning algorithm used to process data and enhance accuracy in image recognition and text extraction.

Connectionist Temporal Classification (CTC) – An algorithm used to train deep neural networks for tasks like speech recognition and handwriting recognition.

Fully Convolutional Network (FCN) – A type of neural network used for tasks such as semantic segmentation, where the goal is to classify each pixel in an image. VisionAid uses FCNs to enhance document layout analysis and text line detection, improving the overall text recognition process.

Glaucoma – A group of eye diseases that can cause vision loss and blindness by damaging the optic nerve. VisionAid aims to assist these individuals by providing auditory output from text and natural scenes.

Maculopathy – A disease related to the central part of the retina, leading to vision loss in the central part of the eye. VisionAid aims to assist these individuals by providing auditory output from text and natural scenes.

Naïve Bayes – A classification algorithm used in related studies to classify data and compare it with other datasets.

Optical Character Recognition (OCR) – A technology that transforms pictures into text. VisionAid employs OCR to recognize and extract text from images captured in real-time.

Region-based Convolutional Neural Network (R-CNN) – A type of deep learning algorithm used for object detection in computer vision.

Scene Image-Text Matching (SITM) – The detection and recognition of scene text from a camera, used for interpreting the context in which text appears within an image.

Support Vector Machine (SVM) – A supervised learning model used for classification and regression analysis. In the paper, it is used to classify and recognize text characters.

Text-to-Speech (TTS) – A technology that transforms text into spoken words. VisionAid uses TTS to provide auditory feedback to visually impaired individuals, enabling them to hear text from both documents and natural scenes.

You Only Look Once (YOLO) – A real-time object detection algorithm that identifies specific objects in videos, live feeds, or images. It is applied in the study for better results and accurate understanding of the model from document reading.

Chapter 2

REVIEW OF RELATED LITERATURE AND STUDIES

It is known for a fact that existing character recognition had already existed and had already been improved. With this segment, the review of the following literature will be done on this chapter, and a more thorough understanding of the studies will be represented from each of the segments. Finally, the application of these models and algorithms will be the basis of the model that researchers plan to build and implement.

Image Character Recognition using Convolutional Neural Networks.

The study of Narayan, A., & Raja M. (2021)Convolutional Neural Network (CNN) is applied within this study for the better text recognition of the following study. This enhances and does preprocessing of the text from images that had been taken, with the preprocessing applied it had the text recognition increased up to 97.59% with a minimal loss of 6.6% thus concluding the application of the CNN can be applied for the further improvement of the text to speech however further improvements are needed for text that are out of the scenery, thus having trouble for the model to recognize signs that are from a far are the text the authors had problem dealing with as the said concern requires a lot of cleaning.

Classification of Documents Extracted from Images with Optical Character Recognition Methods.

Application of OCR (Optical Character Recognition) method is the utilization of text recognition with machine learning methods so that the text recognition will be able to adapt to different types of handwritten and printed documents as introduced by Aydın, Ö. (2021). in this research, The researchers had used the OCR and applied the Naive Bayes Algorithm to class the the data that the system had been given however when compared to the MODI (Microsoft Office Document Imaging Library) was used, from the given example the OCR system detected 346 characters and 61 words from a sample text, while MODI detected 351 characters and 62 words. The word match rates were approximately 85.24% for MODI and 88.52% for the OCR method, with both methods achieving a 97.7% character match rate. Finally suggested in order to have higher accuracy for the application of the OCR, it is recommended by the authors to have much more cleaning data to be applied for the images in order to have a much more clear text recognition on the hand written documents, suggesting to use neural networks for the said statement.

An end-to-end Optical Character Recognition approach for ultra-low-resolution printed text images.

This study from Gilbey, J. D., & Schönlieb, C.-B. (2021), approaches OCR (Optical Character Recognition) on ultra-low-resolution images (60 and 75 dpi). Traditional OCR models struggle with such low resolutions, as demonstrated by the difficulty in reading enlarged low-resolution text images compared to higher resolution ones. This challenge is tackled by a new technique inspired by human vision. Multiple methods had been used in order for the cleaning to be applied, such as Nearest-neighbor with application of interpolation followed by Gaussian filtering with different standard deviation, Blurring the images was used to upscale with the application of Gaussian filters to reduce high-frequency noise making edges less distinct and text easier to recognize.

Accuracy metrics that was accepted were the following Character Level Accuracy, Word Level Accuracy which for 60 dpi images: Character error rate reduced by 64% and word error rate reduced by 73% rate and the 75 dpi images had Character error rate reduced by 35% and word error rate reduced by 51%

The proposed methods include upscaling with various interpolation techniques, modifying the Tesseract's pipeline, and using an ensemble approach to significantly improve the OCR performance on low resolution images. Further research can be suggested to focus on the fine-tuning for specific text genres and optimizing ensemble methods for even better research.

CNN-RNN BASED HANDWRITTEN TEXT RECOGNITION.

According to Hemanth, Jayasree, Venii, Akshaya, and Saranya (2021), the Handwritten Text Recognition (HTR) system delves into text recognition methods aimed at digitizing both handwritten and machine-generated fonts to address challenges such as varying styles, fonts, and character distortions. The datasets used for training and validation are from IAM Dataset consisting of 100,000 images of handwritten text. In their study, images were changed to grayscale, with boundary boxes drawn for line segmentation and word segmentation techniques applied to break down text lines into individual words. The RNN then extracted relevant sequential information. The application of CTC (Connectionist Temporal Classification) significantly improved accuracy, with the proposed model achieving a word recognition accuracy of 98%, which is notably higher than current benchmarks. The researchers suggest that future work should incorporate hybrid datasets to further address challenges such as broken text recognition and many more.

Attention-Based CNN-RNN Arabic Text Recognition from Natural Scene Images.

According to Butt, Muhammad R. R., Muhammad J. R., Muhammad J. A., and Haris (2021), the application of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for Arabic text recognition from natural scene images utilizes deep learning algorithms to read Arabic text, which is often considered complex due to its unique writing style and format. The researchers employed these neural networks to develop a system capable of reading text while eliminating background scenery, color, and other distractions. Results indicate successful recognition at the character, word, and line levels, surpassing existing methods. Concluded by the study that this also outperforms other models like Deep Belief Networks and Multidimensional Long Short-Term Memory Networks, achieving higher character recognition rate. Sequence learning techniques enable direct transcription of images, with contextual modeling possible in both forward and backward directions. From the Arabic OCR engine fine reader used for Arabic Character recognition this OCR obtains a 82.4% to 83.26% character recognition rate however the proposed model by the researchers had a model which obtains 98.73%. Given the availability of effective feature learning algorithms, the research community should shift focus towards tackling more challenging tasks such as natural scene recognition.

Optical character recognition system for Baybayin scripts using support vector machine.

From the study of Pino, Mendoza, and Sambayan (2021), a character recognition system was developed for Baybayin scripts using SVM (Support Vector Machine) to recognize both Baybayin and Latin scripts. The preprocessing step converts the original image into binary data. If the image contains only one component, it proceeds with the algorithm; if it includes an accent, the main component and the accent are separated and processed individually. The SVM is then utilized to classify the models, and character recognition is used to classify the characters based on the extracted features. The authors conducted tests on 1,100 randomly chosen images, from the datasets collected from 9,000+ images for Baybayin characters which were taken from the dataset provided by Nogra (2019) in Kaggle. Achieving the following performance metrics: Accuracy of 98.41%, Precision of 98.68%, Recall of 98.45%, and an F1 Score of 98.57%. The proposed OCR system leveraging SVM models demonstrated high performance in both binary and multiclass classification tasks for Baybayin and Latin scripts, indicating a strong capability to accurately classify characters from these scripts. This system holds potential for future research and improvements in script recognition technologies.

Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection.

In their 2023 study, Fateh, Fateh, and Abolghasemi investigate improvements to Optical Character Recognition (OCR) by application of deep learning models like YOLOv3, SSD, Faster R-CNN, and Layout Parser. These models help to classify and extract textual and non-textual regions within documents, providing coordinates and classifications that enhance OCR readability on the documents. A voting system is applied to determine whether a region is textual or non-textual based on the predictions of the said models.

The authors also implemented Text Line Detection (TLD), which optimizes OCR by performing tasks such as angle correction, spacing adjustment, size normalization, and curvature analysis. Preprocessing steps are applied to distinguish between text and non-text content, followed by the implementation of Tesseract-OCR to convert the processed content into more readable formats.

Results Summary**:** The study showed that applying Text Line Detection (TLD) significantly reduced OCR error rates across various datasets. Specifically, the total error rate was reduced from:

- **6.24% to 3.43%** in the ION Dataset,
- **13.88% to 1.52%** in the Arabic Dataset,
- **6.22% to 3.88%** in the Synthetic Dataset.

These improvements demonstrate the effectiveness of TLD in enhancing OCR accuracy, particularly for complex scripts.

DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts.

From the study of Zhu, Sokhandan, Yang, Martin, and Sathyanarayana (2022), the proposed study on a multi-stage OCR solution for documents with complex layouts aims to enhance OCR performance in complex documents by utilizing a specific sequence of processing steps. The FCN (Fully Convolutional Network) model is employed to perform pixel classification, identifying various segments within the documents, including feature extraction, down-sampling, upscaling, refinement, and classification.

Thus concludes the following:

1. SETR-Heuristic provides the best results for text segments with high accuracy and precision.
2. SepDet-Geometric delivers the best sequential ordering of text.

3. Mask R-CNN and SETR models are effective for detailed layout segmentation, with the Mask R-CNN R50-FPN model being particularly suited for high-accuracy segment classification.
4. For applications requiring real-time processing, Mask R-CNN with R50-DC5 backbone is preferred due to its lower inference time.

This study proves that different models excel depending on the specific application requirements, such as accuracy, inference time, and sensitivity to layout classification, making them usable for complex text documents like newspapers.

Multilingual Text & Handwritten Digit Recognition and Conversion of Regional languages into Universal Language Using Neural Networks.

The utilization of the following Convolutional Neural Network (CNN) for the application into OCR to accurately identify and translate handwritten text and printed text documents from various regional languages into a universal spoken language which is English. On the topic of *Multilingual Text & Handwritten Digit Recognition and Conversion of Regional languages into Universal Language Using Neural Networks* The researchers had prepared preprocessing on the images to cutting importance and relevance to the image, gray scale conversion and inverting images, included segmentation is removing borders and splitting text into characters, datasets that had been used for the data preprocessing and training are the MNIST dataset from the Keras library which contains handwritten digits images . With the model that had been applied the recognition accuracy had around 99% accuracy on the MNIST dataset, showing excellent accuracy on handwritten digits. With this result this opens up for a practical application in future programs that will have number recognition.

Baybayin Script word Recognition and Transliteration Using a Convolutional Neural Network

According to Vilvara, R. A., Hammond, D. S. C., Santos, F. M., & Alar, H. S. (2022), The famous writing script of the Philippines was being revived at the time of the making of the following study for cultural preservation and education purposes, as a growing need of tools for potential improvements of learning of the ancient script, the researchers of this study had suggested the use of the Convolutional Neural Network (CNN) based transliteration model capable of converting Baybayin symbols into corresponding Filipino words. The datasets that were used came from 3 sources as indicated within the study, all of these datasets were tailored for character classification, word detection and transliteration. The baybayin character classification that was proposed model of **CNN-Based Transliteration Model** achieved accuracies of 97.4% and 97.26% from VGG16 based classification, baybayin word detection achieved 96.15%, and the transliteration demonstrated an accuracy of 91.54% on the suggested model. Finally the researchers concluded that the proposed model did achieve promising results, however allow room for improvement as other existing models such as from such as Pino et al.'s system using Support Vector Machines (SVM) with 97.6% accuracy, the use of Levenshtein distance still proves effective for single-word recognition., finally the researchers of this study hopes to contribute further development of Baybayin recognition and transliteration models in the field of computer vision and cultural preservation.

Reading order detection on handwritten documents

Determining the correct reading order for handwritten documents remains a significant challenge due to their inconsistent layouts. Quirós and Vidal (2022) proposed a learning-based approach to address this issue, utilizing a Multilayer Perceptron (MLP) to predict pairwise order relations between text elements. The study introduced two decoding algorithms, Greedy and First Decide Then Decode (FDTD), to arrange text lines or regions in the correct order. Tested on datasets like **OHG**, **FCR**, and **ABP**, the method outperformed traditional rule-based approaches (e.g., top-to-bottom, left-to-right), especially for homogeneous datasets. However, challenges persisted with highly heterogeneous datasets, highlighting the need for advanced classifiers and features. The study also demonstrated that hierarchical processing, sorting regions first and then lines, improved accuracy and efficiency.

This research significantly advances **Handwritten Text Recognition (HTR)** and **Document Layout Analysis (DLA)** by offering a scalable solution for reading order detection.

Conceptual Model of the Study

With the following studies mentioned above, the program implies to gather information and collect data from existing structures and have neural networks applied to the OCR to be developed which is the developmental of the VisionAid that helps secure the accuracy recognition of the text with the application of text recognition, handwritten documents, printed documents, segmentation of the text, ordering, scenery challenges, distance and even the sequence of the complicated text that has different types of writing methods such as arabic, persian and the application of the program towards baybayin writings and the known english language, and finally to implement to Filipino text reader to develop the model.

The research acknowledges and will be utilizing following data flow with agile application and with application of machine learning algorithms studies.

The application of convolutional neural network (CNN) is prevalent from all of the following studies with the application of the support vector machine (SVM) for recognition of baybayin writings, YOLOv9 with application of the improvement suggested by other studies as enhancing it from different sceneries and segmentation are applied for better result and accurate understanding of the model from the document it is reading.

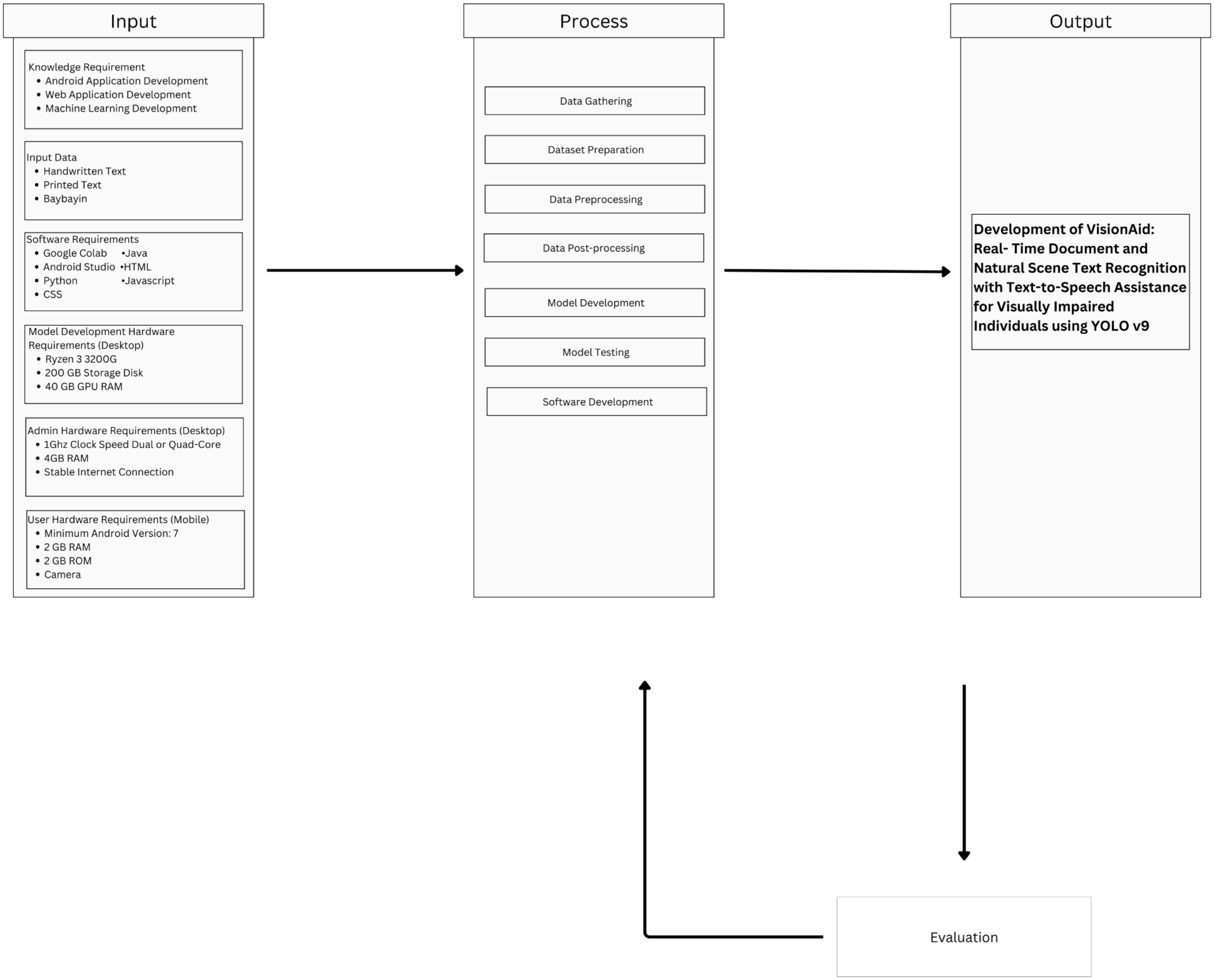

Input
Knowledge Requirement
• Android Application Development
• Web Application Development
• Machine Learning Development
Input Data
• Handwritten Text
• Printed Text
• Baybayin
Software Requirements
• Google Colab •Java
• Android Studio •HTML
• Python •Javascript
• CSS
Model Development Hardware Requirements (Desktop)
• Ryzen 3 3200G
• 200 GB Storage Disk
• 40 GB GPU RAM
Admin Hardware Requirements (Desktop)
• 1Ghz Clock Speed Dual or Quad-Core
• 4GB RAM
• Stable Internet Connection
User Hardware Requirements (Mobile)
• Minimum Android Version: 7
• 2 GB RAM
• 2 GB ROM
• Camera
Process
Data Gathering
Dataset Preparation
Data Preprocessing
Data Post-processing
Model Development
Model Testing
Software Development
Output
Development of VisionAid: Real- Time Document and Natural Scene Text Recognition with Text-to-Speech Assistance for Visually Impaired Individuals using YOLO v9
Evaluation

Figure 1. Conceptual Framework of VisionAID

This conceptual framework provides an overview of the development process for "VisionAid: Real-Time Document and Natural Scene Text Recognition with Text-to-Speech Assistance for Visually Impaired Individuals using YOLO v9." The framework is divided into three main components: **Input**, **Process**, and **Output**, with a feedback loop from **Evaluation** to **Process**.

Input

- **Knowledge Requirements**: The listed requirements include the understanding of the programming languages and expertise on the said field.
- **Input Data**: This specifies the types of data the system will process, including handwritten text, printed text, and Baybayin scripts. The system should be able to recognize diverse text types for accurate assistance.
- **Software Requirements**: The development environment and tools, such as Google Colab, Android Studio, and various programming languages that are mentioned such as Java, Python, HTML, CSS and JavaScript. These are needed to build and test the application.
- **Model Development Hardware Requirements (Desktop)**: This specifies the computing power required for model training, listing hardware components such as a Ryzen 3 3200G processor, 20 GB of storage, and a 40 GB GPU RAM for efficient processing.

- **Admin Hardware Requirements (Desktop)**: Additional hardware requirements for the development environment, such as a 4 GHz dual or quad-core processor, 4 GB RAM, and stable internet connection, are listed to ensure smooth model training and testing.
- **User Hardware Requirements (Mobile)**: The minimum specifications for mobile devices to run the application include Android version 7, 2 GB RAM, 2 GB ROM, and a camera, ensuring compatibility for end users.

Process

- **Data Gathering**: Collecting relevant data (e.g., images of text in different formats, including printed, handwritten, and Baybayin).
- **Dataset Preparation**: Preparing the collected data for machine learning by organizing and labeling it appropriately.
- **Data Preprocessing**: Cleaning and standardizing data for optimal performance, such as resizing images, normalizing, and removing noise.
- **Data Post-Processing**: Refining the outputs of data preprocessing for better model usability (e.g., augmenting datasets or correcting errors).
- **Model Development**: Using YOLOv9 for layout analysis, segmentation, and text recognition.
- **Model Testing**: Evaluating the trained model using key metrics like accuracy and speed.
- **Software Development**: The final application is developed to integrate the model, allowing users to access text-to-speech assistance on their mobile devices. This involves creating a user interface and ensuring that the application is responsive and accessible.

Evaluation

- After completing the development process, the application is evaluated to ensure it meets the desired standards of accuracy, usability, and performance. This involves assessing the model's performance in real-world scenarios to confirm it provides reliable assistance to visually impaired individuals.
- **Feedback Loop**: Based on the evaluation results, adjustments might be necessary. The feedback loop from "Evaluation" back to "Process" (specifically to "Model Development" or "Model Testing") allows for iterative improvements. This loop signifies that, if the application does not meet expectations, the model can be refined, retested, and redeployed until it meets the required standards.

Output

- The final output is the "Development of VisionAid," an application that offers real-time document and natural scene text recognition with text-to-speech assistance. The application is specifically designed for visually impaired individuals, using YOLO v9 technology to accurately recognize text in various environments and provide audio feedback.

This framework illustrates the systematic approach to developing VisionAid, showing how each component contributes to building a robust, user-friendly tool for visually impaired individuals. The inclusion of a feedback loop underscores a commitment to quality and continual improvement.

Table 1. Benchmarking Analysis

Authors	Title	Problems	Algorithm Used	Findings and Conclusion
Narayan, A., & Raja M. (2021)	Image Character Recognition using Convolutional Neural Networks	The objective of this study is to have another method to improve the accuracy of the Optical Character Recognition	Convolutional Neural Network	The returned result had shown great accuracy of 97.59 which only has a 6.6% loss. This suggests that the model is best fit and performed well, however for future improvements, it has challenges on text that are not in good condition for the model to recognize.
Aydın, Ö. (2021)	Classification of Documents Extracted from Images with Optical Character Recognition Methods.	This study aims to apply Naive Bayes algorithm for the subsequent recognition for both handwritten and printed documents that are to be extracted.	Naive Bayes	Both the existing and proposed model had achieve high accuracy from character match rates are approximately 97.4% however the OCR method had a slightly higher word rate of 88.52% compare to the MODI counterpart which is the 85.24% With the text classification of Naive Bayes, the algorithm help resulted in approximately 53% accuracy. The study concludes by suggesting various methods to improve accuracy, such as enhancing image quality, cleaning noisy pixels, refining blob detection, increasing the training dataset size, and exploring alternative classification algorithms like Neural Networks.
Gilbey, J. D., &	An end-to-end Optical	Ultra-low-resolution text are text	Recurrent Neural	The findings of the study showed significant

Schönlieb, C.-B. (2021)	Character Recognition approach for ultra-low-resolution printed text images.	that are completely not ignorable as these things can occur and having the ability for it to be broken down and predicted by Neural network is something prevalent to the OCR models	Network (RNN) and Convolutional Neural Network (CNN)	improvements in OCR performance on low-resolution images. For 60 dpi images, the character error rate was reduced by 64%, and the word error rate was reduced by 73%. For 75 dpi images, the character error rate was reduced by 35%, and the word error rate was reduced by 51%.
Hemanth, G., Jayasree, M., Venii, S., Akshaya, P., & Saranya, R. (2021)	CNN-RNN BASED HANDWRITTEN TEXT RECOGNITION. ONLINE ICTACT JOURNAL on SOFT COMPUTING,	Handwritten text recognition (HTR) aims to digitize handwritten and system generated fonts to overcome challenges such as diverse styles, fonts and character distortions.	1. Convolutional Neural Network (CNN), 2. Recurrent Neural Network (RNN) with applied Short-Term Memory (STM) model applied 3. Connectionist Temporal Classification (CTC)	The proposed model had an accuracy rate of 98% rate for word recognition surpassing existing models. The study recommends the use of hybrid datasets encompassing diverse writing styles, fonts, and text distortions to further extend the system's capabilities.
Butt, H., Muhammad R. R., Muhammad J. R., Muhammad J. A., & Haris, M. (2021)	Attention-Based CNN-RNN Arabic Text Recognition from Natural Scene Images.	To develop a model that is able to recognize Arabic text using deep learning algorithms such as CNN and RNN capable of accurately recognizing Arabic text from	1. Convolutional Neural Network (CNN), 2. Recurrent Neural Network (RNN)	The Attention-Based CNN -RNN for the Arabic text recognition from natural scenes had a high character recognition rate of 98.73%. The findings suggest effective features as mentioned from the accuracy that can be utilized to learn other sets of text and to achieve high accuracy for doing so.

		natural scene images.		
Pino, R., Mendoza, R., & Sambayan, R.. (2021, February 15).	Optical character recognition system for Baybayin scripts using support vector machines.	This study focuses on the OCR system specifically tailored to understand the Baybayin Scripts, a traditional script used in the Philippines that contains various noises and variations in writing styles.	Support Vector Machine	The proposed SVM model achieved high performance both in binary and multi class classification for baybayin scripts through testings on 1,100 randomly chosen images the accuracy gathered is 98.41% Accuracy. Thus this study suggests a solid text character recognition for the baybayin scripts to be read by OCR.
Fateh, A., Fateh, M., & Abolghasemi V. (2023).	Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection.	This study aims to recognize the regions of text documents and non textual regions within documents while utilizing the OCR system as suggest much more advanced techniques that struggles with accurately identifying text in complex layouts such as newspaper and dictionaries.	1. YOLOv3 2. SSD 3. Faster RCNN 4. Text Line Detection (TLD)	The findings suggest that the integration of deep learning models for region detection and TLD for text preprocessing is a promising approach for enhancing OCR technologies, more information can be found directly on the study and results shown as those that are in bold are greater as this shows lower error rates.
Zhu, W., Sokhandan, N., Yang, G., Martin, S., & Sathyanarayana, S. (2022)	DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts. Proceedings of the ... AAAI	The following study hopes to develop an OCR solution capable of accurately reading and processing documents with	Fully Convolutional Network (FCN). Separator Detector (SepDet)	As shown within the results of the research each model with algorithms had different result that were best on each region and segments as SETR-Heuristic provide best result for text segmentation other are offer best sequential

	Conference on Artificial Intelligence,	intricate layouts, which the goal is to understand the layout of the following documents that are presented to reduce error and improve sequencing of document content	Post Processing Methods such as Geometric Approach and Heuristic Approach	ordering of text, thus the study concludes that different models have distinct strengths depending on specific application requirements, such as accuracy, inference time, and sensitivity to layout classification.
Vidhale, B., Khekare, G., Dhule, C., Chandankhede, P., Titarmare, A., & Tayade, M. (2021)	Multilingual Text & Handwritten Digit Recognition and Conversion of Regional languages into Universal Language Using Neural Networks.	The objective of this study is to develop an OCR system using CNN to accurately identify and translate handwritten text and printed text documents from various regional language into english, which also includes diverse type of language writing converting them to the universally understandable format which is english	Convolutional Neural Network	The results provide high accuracy from the 10,000 images provided from the MNIST dataset which achieved 99% accuracy on the testing data.
Vilvara, R. A., Hammond, D. S. C., Santos, F. M., & Alar, H. S. (2022)	Baybayin script word recognition and transliteration using a convolutional neural network. City	The surge of interest of people to learn the ancient scripts of Filipinos suggested new needs for character recognition in the	Convolutional Neural Network,	The model proposed achieved results of an accuracy of 97.4% for character classification, 96.15% for word detection, and 91.54% for word transliteration. Showing further improvements may be needed as existing models from VGG16 and Levenshtein Distance had

	of Makati, Philippines: University of Makati.	field of computers		achieved greater research is its own areas of character recognition for VGG16 and Levenshtein distance for word detection.
Lorenzo Quirós, Enrique Vidal (2022)	Reading Order Detection on Handwritten Documents	Difficulty in determining the correct reading order for handwritten text, which lacks a consistent layout. Most prior solutions focus on printed text.	Greedy Decoding First Decide Then Decode (FDTD) and Multilayer Perceptron (MLP) .	The proposed approach achieves better results than traditional methods (e.g., TBLR), especially in homogeneous datasets like OHG and FCR. Hierarchical application further improves accuracy. Performance on highly heterogeneous datasets like ABP remains a challenge, suggesting the need for more sophisticated classifiers and features.

This table presents a benchmarking analysis of various studies focused on Optical Character Recognition (OCR) using different algorithms and models. Each study explores a distinct approach to improve the accuracy and functionality of OCR systems, especially for different languages, text types, and document layouts. provided below are brief recommendations of each study for future research.

Recommendations from the researchers on the benchmarking

1. **Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection.**

   Future work should focus on expanding the dataset and exploring other languages with unique script characteristics to develop efficient methods for.

2. **Attention-Based CNN-RNN Arabic Text Recognition from Natural Scene Images.**

   Due to the availability of good feature learning algorithms, the focus of the community should move to harder problems such as natural scene recognition.

3. **CNN-RNN Based Handwritten Text Recognition.**

   In future work, the researchers intended to improve the work by making use of hybrid datasets and experimenting with different activation functions, also increasing the number of neural network layers. Further, we aim to enhance the work by implementing online recognition and extend it to different languages, additionally we can promote the system to recognize degraded text or broken characters.

4. **An end-to-end Optical Character Recognition approach for ultra-low-resolution printed text images.**

   An improvement would be to perform the page segmentation using just one upscaling, and to use this segmentation for each of the line recognition runs; we did not attempt to implement this in our experiments. We could also use fewer upscalings or different upscalings from the ones we chose: it is unclear how to best choose them, and our choice was a somewhat arbitrary selection of filters for reducing the high frequency components of the images. Finally, it would be very useful to adapt this approach to tackle a challenge such as the ICDAR 2015 Competition on Text Image Super-Resolution [11], which used very different types of image from the ones we studied in this work.

5. **DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts.**

Future works should integrate some type of segmentation in order for the system to have better recognition of manuscripts in order for it to be much more complex text reading not only for words but also for segments on a specific picture. thus the following research makes it recommendable and no flaws

6. **Classification of Documents Extracted from Images with Optical Character Recognition Methods.**

In this study the utilization of OCR against Microsoft Office Document Imaging Library (MODI), results that the application of the OCR is better for letter and word recognition, better than the said model, however with the results future application of neural network is recommended.

7. **Image Character Recognition using Convolutional Neural Networks.**

The application of the CNN is formidable in the section of OCR thus highly recommended by the researchers as recommended by Narayan, A., & Raja M. (2021) showing high performance as shown on the Table 1 of the high accuracy in recognition, however room for improvement in the section of the natural scene and handwritten had given limited success due to the difficulty in contour based techniques.

8. **Multilingual Text & Handwritten Digit Recognition and Conversion of Regional languages into Universal Language Using Neural Networks.**

   With application of CNN, greater heights had been achieved as proven by the study from Vidhale, B.et al (2021), better understanding of numbers from different languages converted to universal language is achieved with the application of the CNN and with data the researchers had gathered, however complex mathematical signs is something can be studied in the future for potential mathematical equation understanding.

9. **Optical character recognition system for Baybayin scripts using support vector machine.**

   On the study of Pino, R. et al (2021). The utilization of SVM for character recognition for the baybayin scripts shows great results for the understanding of the “Kudlit” as this can be tricky to be understood. Lastly further studies can be looked into further machine learning algorithms for better classification accuracy.

10. **Baybayin Script Word Recognition And Transliteration Using A Convolutional Neural Network.**

    From the study of Vilvara, R et al (2022), high accuracies had been shown for the application of the CNN. further improvements can be applied by increasing the database will result in improvements on the transliteration algorithm.

11. **Reading order detection on handwritten documents**

The study of Quirós, L., & Vidal, E. (2022), highlights the advances of handwritten text to improve the DLA. Further improvements and gaps that the researchers acknowledged is the reading of old handwritten documents and harsh sceneries.

Synthesis

The developmental of the new technologies nowadays cannot be ignored, as new improving technological advancement are always being offered to the field of research such as the application of Optical Character Recognition (OCR) which is proven reliable by Aydın, Ö. (2021) to solve the following problems and apply neural network for better character recognition, while the utilization of Google's Tesseract Model is offered in multiple studies shown which had been from the previous studies, the researchers had worked on to apply CNN as suggested by Narayan, A., & Raja M. (2021). to the model, despite proven in other studies such as Pino, R (2021), higher accuracy with the utilization of SVM showed great results, the use of CNN is proven reliable throughout all of the existing studies. The researchers plan to create a database to expand the datasets where the model can train to address multiple studies such as from Vilvara, R et al (2022) , Fateh, A et al (2023) that bigger databases can increase the transliteration and accuracy of the model of CNN to increase the accuracy. The application of many line segmentation and document layout analysis (DLA) is highly suggested from the research of Zhu, W. (2022) as this is an improvement acknowledged by multiple researchers such as from the study of Butt, H. (2023). Finally acknowledging the possible improvements on Reading Order Detection by Quirós, L., & Vidal, E. (2022 for the improvement of the DLA as a final touch of the program.

The following it the outline objective of the group to build in order to createVisionAID

1. Optical Character Recognition (OCR):

   Implementing advanced OCR technology to enable accurate text extraction from various sources.

2. Machine Learning with Convolutional Neural Networks (CNN):

   Utilizing CNNs to enhance image recognition and improve the overall performance of the application.

3. Dynamic Layout Analysis (DLA):

   Applying DLA to optimize program structure and interface design for efficient navigation and interaction.

4. Comprehensive Dataset:

   Leveraging a large and diverse dataset to train the models for better accuracy and adaptability to different real-world scenarios.

5. Reading Order Detection:

   Incorporating algorithms to determine and present text in the correct reading order, ensuring accessibility and usability.

The development of VisionAid is guided by the recommendations and gaps identified in previous research. Following the Agile methodology, the team aims to deliver an application that is both highly functional and tailored to the needs of its target audience.

This iterative approach ensures continuous improvement, user feedback integration, and alignment with the usability expectations of individuals with visual impairments.

Chapter 3

DESIGN AND METHODOLOGY

This chapter presents the methodology of the study. This part elaborates the different approaches of methods to prove the efficiency of the algorithm. Contents in this chapter states on how the researchers' approach in terms of research and development to the study.

Research Design

This study will follow the mixed-method research design, integrating both experimental and descriptive research design to develop the model by incorporating the chosen algorithm, and evaluate the application.

According to the definition given by Saigo(2023) of experimental research design, a research approach used to investigate the interaction of dependent and independent variables, which can be utilized to determine a cause and effect relationship. This method allows the researchers to control and manipulate variables to observe their effects. In this study, the researchers modified and customized the YOLO v9 object detection model by incorporating Document Layout Analysis at the initial stage of building the OCR model. By integrating layout analysis, the proposed system is better equipped to identify and differentiate between various elements of a document, whether it is handwritten or printed. The proponents further adapted YOLO v9 to recognize and classify components like text blocks, and tables, which improves the model's ability to understand document structure.

While descriptive research, as defined by Hassan(2024), it aims to describe or record the traits, behavior, attitudes, opinions, or perceptions of a group or population under study, which is useful for VisionAID to gather insights regarding the visually impaired individuals who use the application.

Research Methodology

This study focuses on developing VisionAID, a robust pipeline designed for document layout analysis, word segmentation, and optical character recognition (OCR). The research utilizes a combination of deep learning techniques, using YOLOv9 to optimize real-time text recognition and document analysis.

Instruments and techniques

Data Gathering Procedure

Datasets from Publaynet, Common Object in Context Text(COCO-Text), TextOCR, Facebook Research's IMGUR5K, and Kaggle will be collected and used to train separate models tailored to specific tasks. Each dataset will be used independently, with Publaynet focusing on Document Layout Analysis, COCO-TEXT on Natural Scene Text recognition, and Facebook Research's IMGUR5K on Optical Character Recognition. These datasets include machine-printed text, handwritten text, natural scene text, and document layout data, all of which are essential in building VisionAID's pipeline.This approach ensures that different models are optimized for their respective tasks, enhancing VisionAID's performance in object detection, text recognition, and document analysis.

Data Preparation

Because of the varying formats used by these datasets with most of them following the COCO data format, the researchers will develop a program to standardize them into a unified format compatible with YOLO training. This process involves converting existing annotations to YOLO's format, creating a dedicated Yet Another Markup Language (YAML) file.

Pre Processing

During the preprocessing stage, images are enlarged for better quality and to get the best input for the other phases of VisionAID prediction for best results.

Model Development

A. Algorithm Design and Model Architecture

In this study, the researchers will be utilizing a deep learning framework, specifically YOLOv9, as well as different algorithms in order to develop the pipeline. The following algorithms are as follows:

i. Document Layout Analysis

In order for our pipeline to have a better understanding of the image or document that it is scanning, a document layout analysis is to recognize all document component (Text, Title, List, Table, and Figure), Papers with Code (2022) defined document layout analysis (DLA), as the process of determining the physical structure of a document or document components.

ii. Optical Character Recognition

According to Amazon Web Services (2022) Optical Character Recognition (OCR), refers to the process of converting an image that contains text into a machine readable text format. Google Tesseract Engine will be used to extract text, which is essential in our pipeline as it enables VisionAID to recognize, analyze, and extract text from an image allowing users to read, and navigate text prompts on the application for better information extraction.

The Document Layout Analysis Model will be built and trained using the YOLO architecture, specifically the *YOLO v9* Architecture will be utilized by VisionAID, to consider the computational performance of this architecture in order to give real time detection. According to Wang, Yeh, and Liao (2024), compared to its predecessor, YOLOv9 is a key step forward in real-time object detection, bringing notable enhancements in efficiency, precision, and flexibility.

B. Pipeline Development

The pipeline will be developed to the researcher's specifications, using python as a programming language, google colab as its environment. Model and application development is done during this phase. The development will be done to the researcher's specifications.

C. Post Postprocessing

The Document Layout Analysis (DLA) in the post-processing phase defines the typical flow of document content to be read, cuts out pieces such as text or titles of interest, and then enlarges them in order to read the contents with even higher detail. These refined regions are then passed to the OCR engine in order to obtain a very good text extraction.

D. Model Testing

In evaluating the performance of VisionAID's machine learning model, the following metrics, as outlined by Ultralytics (2023), will be used to assess both the model's effectiveness and overall performance:

i. Precision

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Equation 1. Precision

Precision measures the accuracy of the positive predictions made by the model. It is defined as the number of true positives divided by total predicted positives.

ii. Recall

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Equation 2. Recall

Recall, also known as *sensitivity,* measures the ability of the model to identify all relevant instances. It is defined as the number of true positives divided by the actual positive.

iii. F1 Score

$$F1\ Score\ = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Equation 3. F1 Score

F1 Score is defined as the harmonic average between precision and recall.

iv. Mean Average Precision

$$mAP\ = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

Equation 4. Mean Average Precision

Mean Average Precision measures the model's overall accuracy in detecting objects.

Software Development Process

The researchers intend to follow the Agile Software Development process when building the application. This approach ensures that the application will be developed iteratively, allowing for flexibility and continuous feedback, which in turn helps to adapt quickly to changes and deliver a high-quality product that meets user needs.

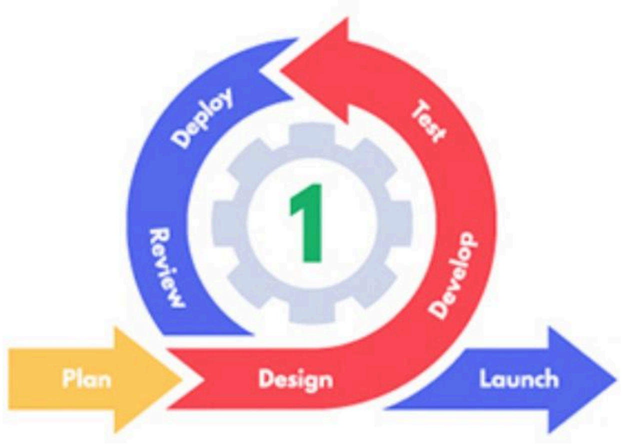


Figure 2. Agile Software Development (Petit, 2021)

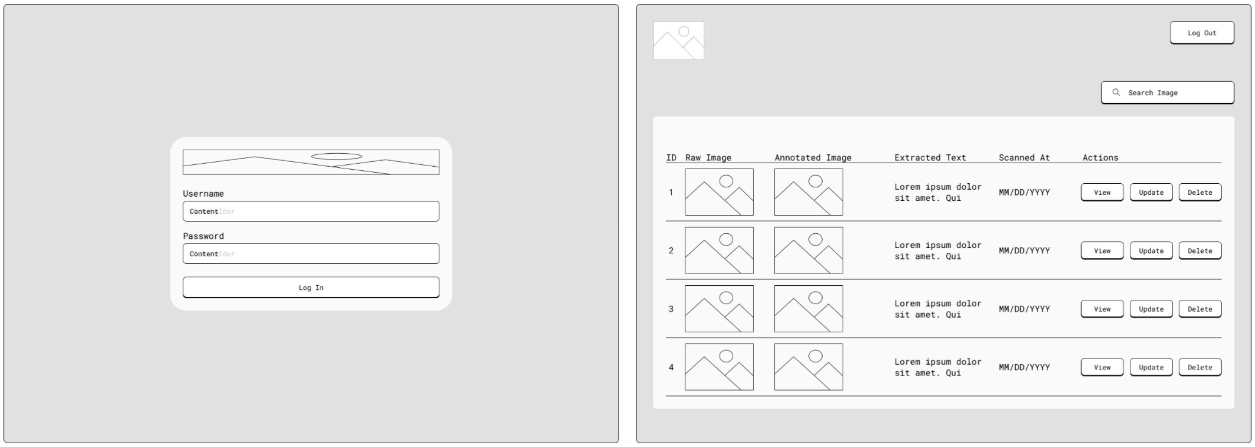
Figure 2 illustrates the developmental stages that the researchers must go through in order to build VisionAID.

Plan

During this phase, the researchers gather knowledge defining the objective and scope of the application. This involves application development, web server development, and system integration. The researchers also select the appropriate algorithm in order to integrate the machine learning pipeline into the application, this includes other necessary libraries.

Design

In this phase, the application's user interface and experience are conceptualized through wireframes and prototypes. This helps to outline the application's layout, navigation flow, and interactive elements, ensuring it is user friendly and accessible to visually impaired users. The researchers also define the system architecture, planning how different components such as the model, user interface, and backend servers will integrate and communicate.



ID	Raw Image	Annotated Image	Extracted Text	Scanned At	Actions
1			Lorem ipsum dolor sit amet. Qui	MM/DD/YYYY	View Update Delete
2			Lorem ipsum dolor sit amet. Qui	MM/DD/YYYY	View Update Delete
3			Lorem ipsum dolor sit amet. Qui	MM/DD/YYYY	View Update Delete
4			Lorem ipsum dolor sit amet. Qui	MM/DD/YYYY	View Update Delete

Figure 3. Low-Fidelity Wireframe of VisionAID Admin Web Application

Shown in Figure 3. is the Low-Fidelity Wireframe of the VisionAID Admin Web Application, it consists of two screens, the login screen, and the dashboard screen. The login screen, located on the left, provides a simple interface where administrators of the application can enter their credentials and log in to the application. Once logged in, administrators are redirected to the Dashboard screen on the right, which serves as the central hub for managing scanned image data. Administrators can use the dashboard in order to keep track of the scanned images, they have control whether they would like to view/update/delete a record in the database, these scanned images can be used to improve the model by training it with new data.

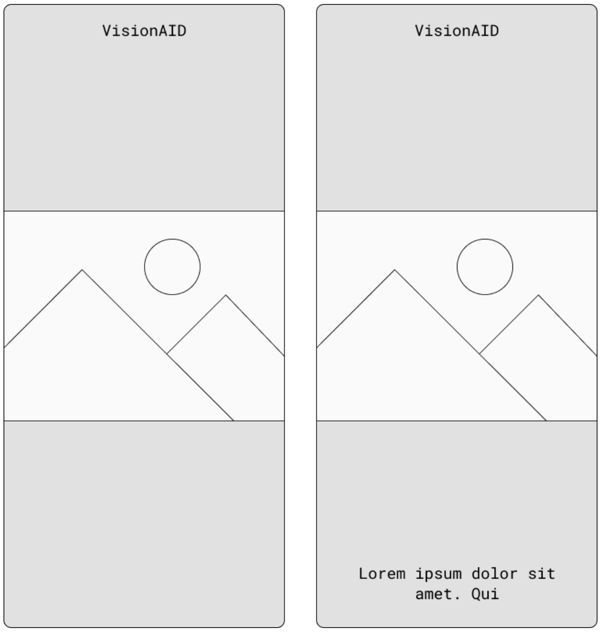


Figure 4. Low-Fidelity Wireframe of VisionAID User Mobile Application

Figure 4. shows the low-fidelity wireframe of VisionAID application designed for the client or end user. The left panel displays a screen where the user can utilize their camera to scan an image. Meanwhile, the right panel shows the text that has been scanned or extracted from the image, which is then read back to the user.

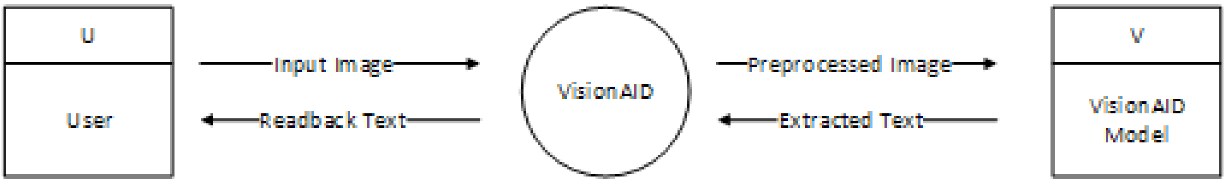


Figure 5. VisionAID Data Flow Diagram Level 0

Figure 5. shows the overview of the system through a Data Flow Diagram Level 0. It shows the interaction between the user and the model through the VisionAID system, where the initial input starts from the user scanning images and the system preprocessing it before analyzing using the model for extraction of text and analyzation of document layout, which will then be read back to the user using text to speech.

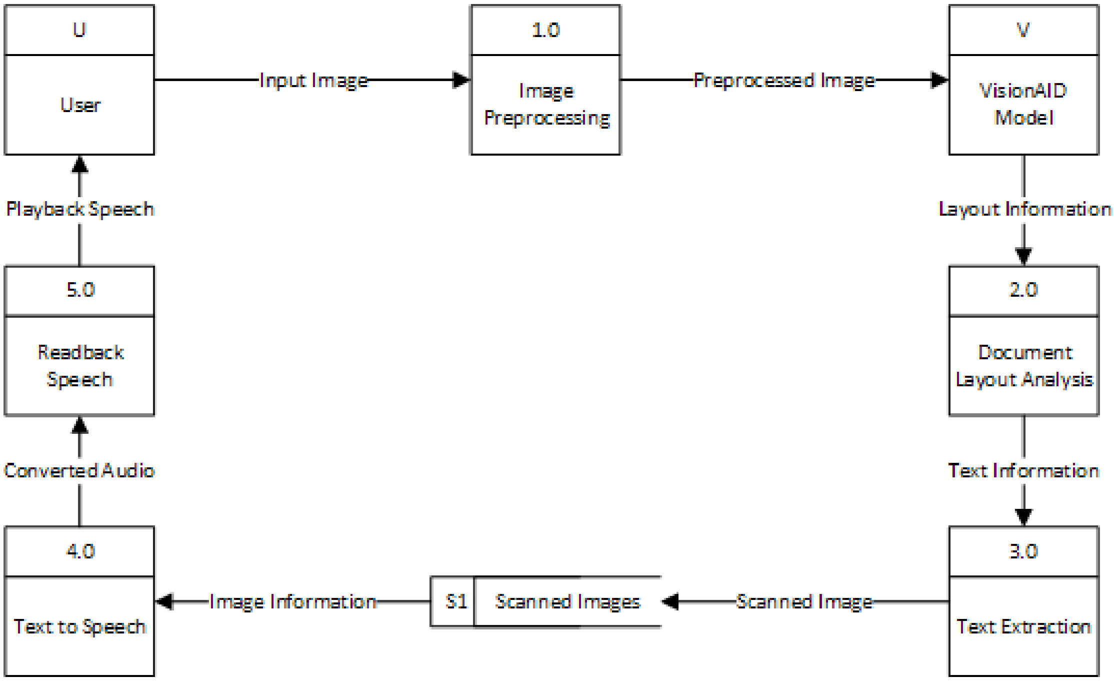


Figure 6. VisionAID Data Flow Diagram Level 1

Figure 6. shows a more detailed view of the VisionAID system, compared to DFD 0, it shows the breakdown of the processing into smaller pieces. Similarly, it starts with the image as an input which will then be processed, and analyzed using the VisionAID system. The scanned image will then be saved into the database for future reference, as well as the information of the scanned image will be converted into an audio using text to speech which will be read back to the user.

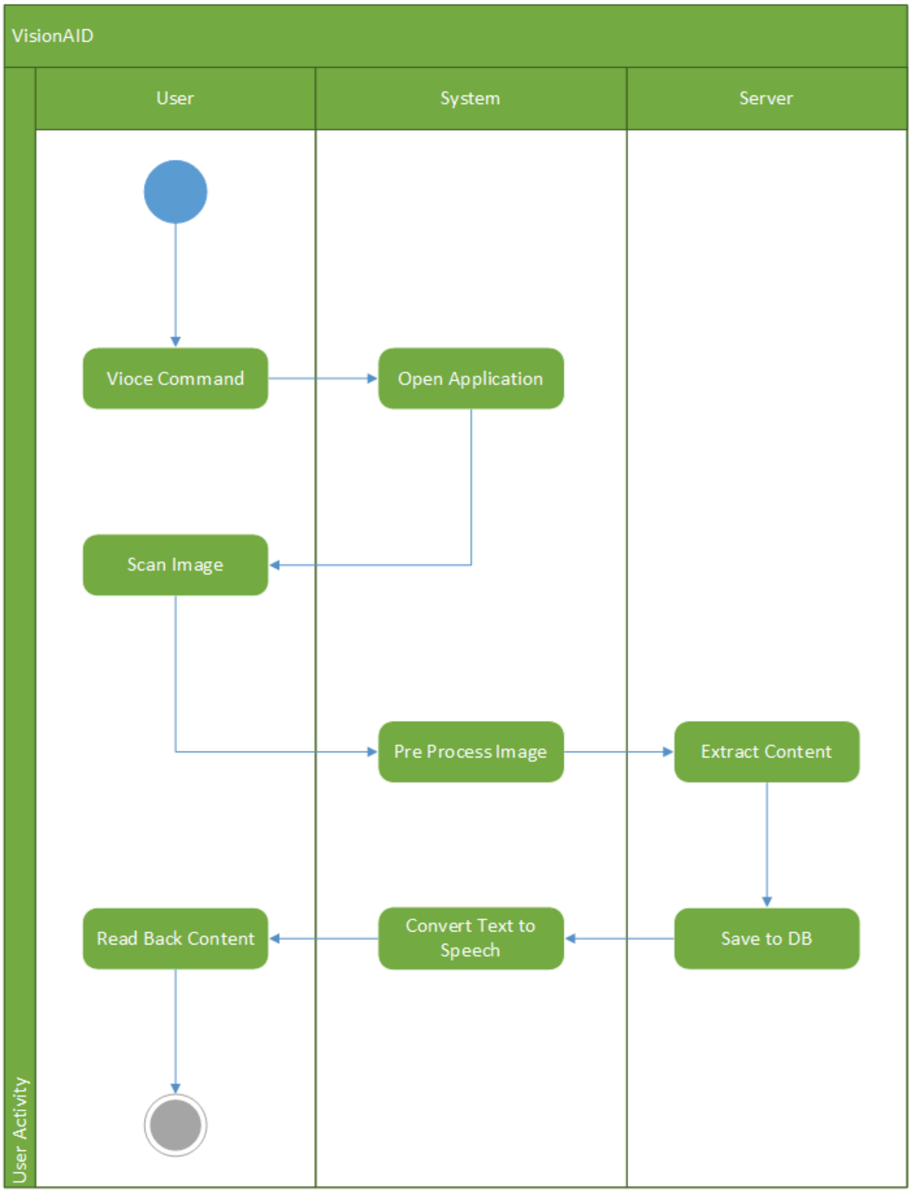


Figure 7. User Activity Diagram

In this Figure 7. shows the sequence of interaction between the end-user, VisionAID application or system, and the Server. Initially the user sends out a voice command in order to open the application, which the system process and opens the VisionAID application ready to scan an image, after the image has been scanned, it will be pre-processed before sending to the server for the utilization of the VisionAID model, and extraction of text and logging of the image in the database. Lastly, the model will then send back the extracted text from the image to the application which will then read back its content to the user.

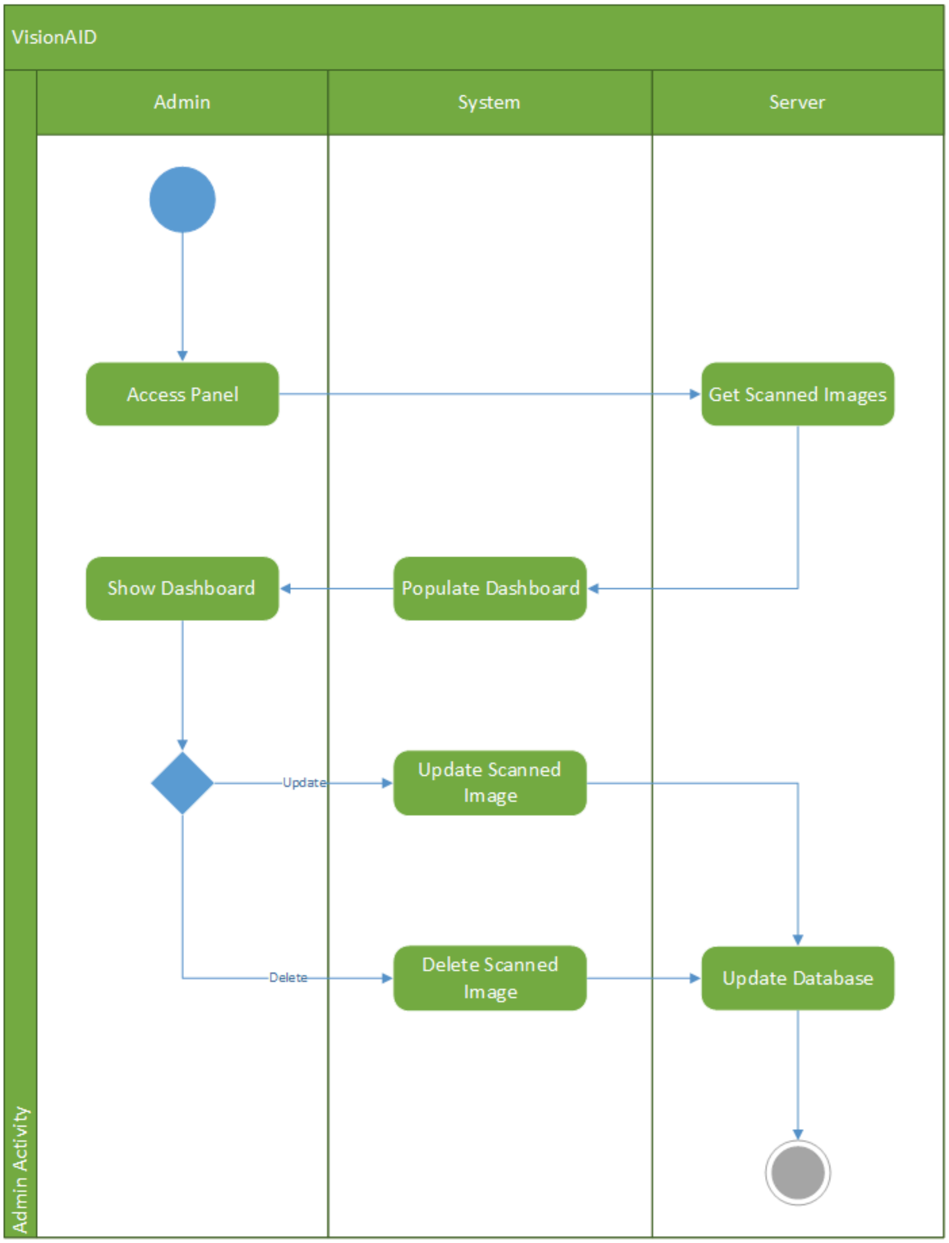


Figure 8. Admin Activity Diagram

An admin activity is shown in Figure 8. starting with accessing of the admin panel, and the VisionAID server will get all data in the database for the scanned images, which will then populate the html layout for the dashboard that will be shown or displayed to the user. The administrator has a choice of whether they will delete or update data which then be processed accordingly and will be updated in the database.

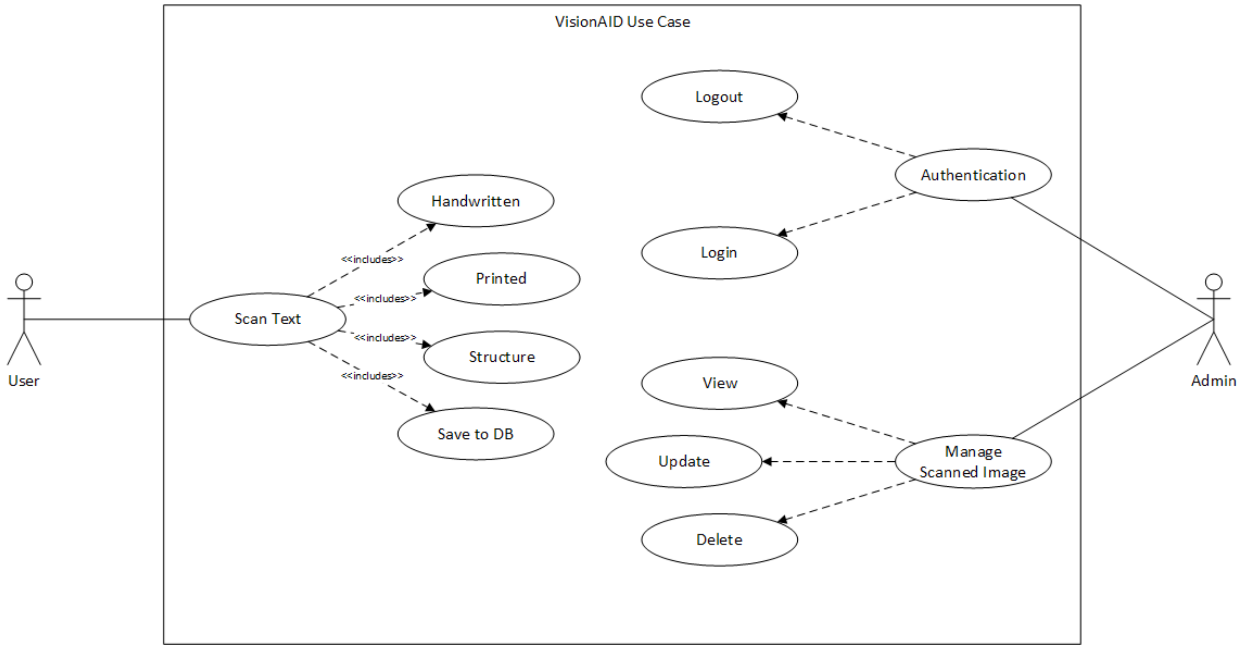


Figure 9. VisionAID Use Case Diagram

Figure 10. depicts a use case diagram that visualizes how different actors interact with the VisionAID application, highlighting key features of the application.

Actors:

- **User** - Visually impaired individuals who use the VisionAID application.
- **Admin** - Overall or system admin of the VisionAID application.

Use Cases:

- **Scan Text**
  - **Handwritten** - Allows the user to scan handwritten documents.

    - **Printed** - Allows the user to scan printed documents.
    - **Structure** - Analyze the structure of the document.
    - **Save to DB** - Allows the user to save the scanned image to the system database.
- **Authentication**
    - **Login** - Admin can enter their credentials to access the system dashboard.
    - **Logout** - Exit out of the application session.
- **Manage Scanned Images**
    - **View** - View all of the scanned image/s from the database.
    - **Update** - Updates the attribute of the scanned image/s from the database.
    - **Delete** - Delete scanned image/s from the database.

Relationships:

- **User** - Has access to the core functionalities of the VisionAID application, allowing them to scan a document and utilize the model to extract and analyze the document structure.
- **Admin** - Has a higher privilege of accessing the application, allowing view/update/delete of data from the database.

Develop

The mobile application will be built using android's native language which is Java, while for the web application, which is for admin to use, HTML, CSS, and JavaScript will be used. This includes the integration of frontend and backend connection, ensuring smooth communication between both ends.

For the backend server, the application will be built using Python Flask, a lightweight web framework designed to be run on a WSGI(Web Server Gateway Interface) server. The backend will operate on an Ubuntu operating system.

The WSGI server that will be utilized in this research is Gunicorn, a python WSGI server for UNIX. which will manage the Flask application's execution and handle HTTP requests.

NGINX will also be utilized as the web server and reverse proxy, on top of Gunicorn, it will handle incoming HTTP traffic, server static files, and route dynamic requests to Gunicorn.

All of these components are connected, ensuring the efficiency of handling web requests and scalability of the applications.

Test

In this phase, the researchers will quantify the results of VisionAID through different test cases to evaluate the functional and non-functional aspects of the application. Some functional tests will include checking the accuracy of text recognition, evaluating the real-time operation of the system, analyzing the layout of the documents to be processed using the system and separation of words contained in the document. Non-functional tests will include testability, capacity, performance, usability, resilience, compatibility, and sustainability to confirm that the system can process real-time data and is compatible with gadgets. Given this comprehensive testing approach, VisionAID stands to provide the best experience to visually impaired users.

Deploy

In this phase, the focus is on ensuring the application is up and running according to the specified requirements, with all intended features implemented and functioning as expected. The researchers will deploy the application using Hostinger's hosting services, which will facilitate the setup of the application server within a Virtual Private Server (VPS).

Review

Lastly, during this phase, the researchers gather feedback from the visually impaired individuals on what is needed to be improved on the system. Also during this phase, the researchers also assess all of the implemented features, and fix any bugs or errors in the system if there's any.

User Satisfaction of the application will be measured using Likert Scale, this involves users rating their satisfaction on a scale, where a higher number indicates greater satisfaction, shown in Table 1.

Table 1. User Satisfaction Score

Score	Interpretation
1	Strongly Dissatisfied
2	Dissatisfied
3	Neutral
4	Satisfied
5	Strongly Satisfied

The average score will be used to quantify the overall user satisfaction of VisionAID as shown in Table 2.

Table 2. Overall User Satisfaction

Range of Mean Overall Score	**Interpretation**
1.50 and below	Strongly Dissatisfied
1.51 - 2.50	Dissatisfied
2.51 - 3.00	Neutral
3.51 - 4.00	Satisfied
4.51 - 5.00	Strongly Satisfied

A. ISO 25010

A standard for software product quality that defines a model for evaluating software. It includes eight quality characteristics such as functionality, reliability, usability, and performance efficiency, to ensure comprehensive assessment of software quality.

B. Technology acceptance model (TAM)

A framework that helps understand how users come to accept and use technology. It is often used in system implementation to analyze user acceptance factors, focusing on two main components:

1. Perceived Usefulness (PU) – how much users believe that using the system will improve their performance or productivity.
2. Perceived Ease of Use (PEOU) – how much users think the system will be easy to use.

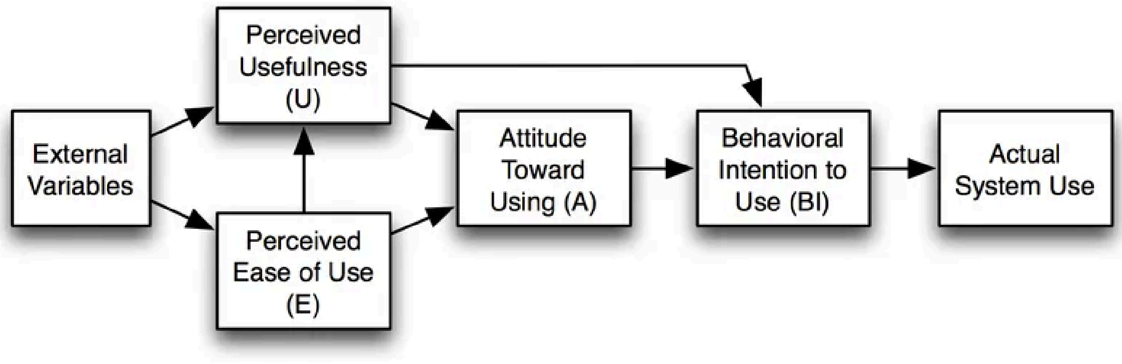

Perceived Usefulness (U)
External Variables
Perceived Ease of Use (E)
Attitude Toward Using (A)
Behavioral Intention to Use (BI)
Actual System Use

Figure 11. Technology Acceptance Model

Following the figure presented, the researchers plan to use the following model to measure the acceptance and the usability of each individual that will be testing the device.

References:

Aydin, O. (2021). Classification of Documents Extracted from Images with Optical Character Recognition Methods. ArXiv.org. https://arxiv.org/abs/2106.11125

Amazon Web Services. (2022). What is OCR? - Optical Character Recognition Explained - AWS. Amazon Web Services Inc. https://aws.amazon.com/what-is/ocr/#:~:text=Optical%20Character%20Recognition%20(OCR)%20is,scan%20as%20an%20image%20file.

Burton, M. J., Ramke, J., Marques, A. P., Bourne, R. R. A., Congdon, N., Jones, I., Ah Tong, B. A. M., Arunga, S., Bachani, D., Bascaran, C., Bastawrous, A., Blanchet, K., Braithwaite, T., Buchan, J. C., Cairns, J., Cama, A., Chagunda, M., Chuluunkhuu, C., Cooper, A., ... Faal, H. B. (2021). The Lancet Global Health Commission on Global Eye Health: vision beyond 2020. The Lancet Global Health, 9(4), e489–e551. https://doi.org/10.1016/s2214-109x(20)30488-5

Butt, H., Muhammad R. R., Muhammad J. R., Muhammad J. A., & Haris, M. (2021). Attention-Based CNN-RNN Arabic Text Recognition from Natural Scene Images. Forecasting, 3(3), 520–540. https://doi.org/10.3390/forecast3030033

Cleveland Clinic. (2020). *20/20 Vision: What It Means & Corrective Methods*. Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/8561-2020-vision

Fateh, A., Fateh, M., & Abolghasemi V. (2023). Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection. Engineering Reports. https://doi.org/10.1002/eng2.12832

Gilbey, J. D., & Schönlieb, C.-B. (2021). An end-to-end Optical Character Recognition approach for ultra-low-resolution printed text images. ArXiv.org. https://arxiv.org/abs/2105.04515

Hassan, M. (2024). Descriptive Research Design – Types, Methods and Examples. Research Method. https://researchmethod.net/descriptive-research-design/

Hemanth, G., Jayasree, M., Venii, S., Akshaya, P., & Saranya, R. (2021). CNN-RNN BASED HANDWRITTEN TEXT RECOGNITION. ONLINE ICTACT JOURNAL on SOFT COMPUTING, 1. https://doi.org/10.21917/ijsc.2021.0351

Mcleod, S. (2023). Likert Scale Questionnaire: Examples & Analysis. https://www.simplypsychology.org/likert-scale.html

Narayan, A., & Raja M. (2021). Image Character Recognition using Convolutional Neural Networks. https://doi.org/10.1109/icbsii51839.2021.9445136

Petit, C. (2021). Agile Software Development: What It Is and the Benefits It Offers. Softrizon. https://www.softrizon.com/blog/agile-software-development/

Papers with Code. (2022). Document Layout Analysis | Papers with Code. Paperswithcode.com. https://paperswithcode.com/task/document-layout-analysis

Philippine Statistics Authority. (2023, December 22). Preliminary 2023 First Semester Official Poverty Statistics. https://www.psa.gov.ph/statistics/poverty

Pino, R., Mendoza, R., & Sambayan, R.. (2021, February 15). Optical character recognition system for Baybayin scripts using support vector machine. ResearchGate; PeerJ. https://www.researchgate.net/publication/349323182_Optical_character_recognition_system_for_Baybayin_scripts_using_support_vector_machine

ROQUE Eye Clinic. (2024). Persons with disabilities. https://www.eye.com.ph/about-us/policies/pwd/

Saigo, H. (2023). Experimental Research Design | Definition, Components & Examples. Study.com. https://study.com/academy/lesson/the-true-experimental-research-design.html

STATISTA. (2024, January). Philippines: major mobile OS by market share 2024. https://www.statista.com/statistics/931129/philippines-mobile-os-share

Tesseract-OCR. (2024). Tesseract. GitHub. https://github.com/tesseract-ocr/tesseract

Tigerschiold, T. (2022). What is Accuracy, Precision, Recall and F1 Score? Labelf.ai. https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score

Ultralytics. (2023). YOLO Performance Metrics. Ultralytics.com. https://docs.ultralytics.com/guides/yolo-performance-metrics/

Vidhale, B., Khekare, G., Dhule, C., Chandankhede, P., Titarmare, A., & Tayade, M. (2021). Multilingual Text & Handwritten Digit Recognition and Conversion of Regional languages into Universal Language Using Neural Networks. https://doi.org/10.1109/i2ct51068.2021.9418106

Villanueva, M. A. (2022). Lost focus in the fight vs blindness. Philstar.com; Philstar.com. https://www.philstar.com/opinion/2022/11/30/2227378/lost-focus-fight-vs-blindness

Vilvar, R. C., Hammond, D. S. C., Santos, F. M. R., Alar, H. S. (2022). Baybayin Script Word Recognition and Transliteration Using a Convolutional Neural Network. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4004853

Wang, C.-Y., Yeh, I-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. https://arxiv.org/pdf/2402.13616

Wei, J., Zhan, H., Tu, X., Lu, Y., & Pal, U. (2023). Scene Text Recognition with Image-Text Matching-guided Dictionary. ArXiv.org. https://arxiv.org/abs/2305.04524

WHO. (2019). World report on vision. https://www.who.int/docs/default-source/documents/publications/world-vision-report-accessible.pdf

Zhu, W., Sokhandan, N., Yang, G., Martin, S., & Sathyanarayana, S. (2022). DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts. ArXiv.org. https://arxiv.org/abs/2202.01414

Vilvara, R. A., Hammond, D. S. C., Santos, F. M., & Alar, H. S. (2022). *Baybayin script word recognition and transliteration using a convolutional neural network*. City of Makati, Philippines: University of Makati. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4004853

Zahra, S. (2023, April 29). OCR Segmentation with Python Code - Sana'a Zahra - Medium. Medium. https://sanaazahra.medium.com/ocr-segmentation-with-python-code-f3251114ee48