BIS INFORMATION TECHNOLOGY AND COMPUTER SCIENCE

Volume 2 2025 | DOI: 10.31603/bistycs.185
6th Borobudur International Symposium on Science and Technology (BIS-STE) 2024

UNIMMA PRESS
UNIVERSITAS MUHAMMADIYAH MAGELANG

Integrating YOLOv8, EasyOCR, and GTTS for text detection in assistive technology for the visually impaired

Maria Bestarina Laili[1*], Kartika[1], Muhammad Zaki Abdulah[1], Syuraih Amiruddin[1], Egi Sunardi[1], Jelita Permatasari[1]

[1] Department of Electrical Engineering, Universitas Singaperbangsa Karawang, Indonesia
[*] Corresponding author: maria.bestarina@ft.unsika.ac.id

Abstract

Technology for visually impaired individuals has advanced, but accessing text-based information remains challenging. Accurate text detection, clear reading, and voice conversion are essential. YOLOv8, EasyOCR, and Google Text-to-Speech (GTTS) are cutting-edge technologies that can be integrated to address this need. This study aims to develop a system combining YOLOv8 for text detection, EasyOCR for text recognition, and GTTS for text-to-speech conversion, focusing on improving accessibility for the visually impaired. The system operates in several stages. First, YOLOv8 detects text in images in real-time. Next, EasyOCR extracts text from the detected regions. Finally, GTTS converts the recognized text into clear speech. A diverse text image dataset was used for training and testing the detection model, while user testing was conducted to assess system usability and effectiveness. The developed system successfully detects and reads text with high accuracy and converts it into clear speech. System evaluation revealed significant improvements in information accessibility for the visually impaired, with users responding positively to its speed, accuracy, and ease of use. Integrating YOLOv8, EasyOCR, and GTTS into a single solution presents an innovative approach to text detection, recognition, and conversion for visually impaired individuals. This system demonstrates significant potential to enhance independence and quality of life by improving access to text-based information. The study contributes to assistive technology development and opens doors for further research into practical applications and system refinement.

Published:
April 28, 2025

Selection and Peer-review under the responsibility of the 6th BIS-STE 2024 Committee

Keywords

Text detection, Assistive technology, Visually impaired

Introduction

The visually impaired rank first among all other disability categories in terms of limiting [1]. This is due to their limitations in vision and mobility. According to estimates from the Ministry of Health of the Republic of Indonesia, the number of visually impaired

individuals in Indonesia accounts for 1.5% of the total population. With Indonesia's current population exceeding 270 million, the number of visually impaired individuals is estimated to be around 4 million. This figure is significant. Addressing this issue, visually impaired individuals require technology to assist them in carrying out daily activities. One of the main technological innovations in this area is the development of assistive tools designed to improve the quality of life for persons with disabilities, especially the visually impaired. Among various assistive technologies, systems that can detect, read, and convert text into speech hold great potential for enhancing information accessibility and independence for the visually impaired.

In technological advancements, various aspects can be utilized to aid users in their daily lives, such as Text-to-Speech (TTS) technology. Text-to-Speech (TTS) is a technology that enables a device to convert written text and images into speech. It is based on two main principles: converting text into phonemes and converting phonemes into speech. First, the text containing sentences is converted into a series of sound codes, typically represented by phoneme codes, duration, and pitch. Then, these codes are processed to generate speech signals that match the input. TTS differs from Speech-to-Text (STT), often referred to as Speech Recognition. Speech Recognition allows computers to receive input in spoken language. Spoken words are converted into digital signals by transforming sound waves into numerical data, matching them with specific codes, and comparing them to stored patterns. The recognized spoken language output can be displayed as text or read by technical equipment. TTS has been widely implemented across various fields, primarily to assist individuals with visual impairments. One commonly used speech engine for TTS implementation is Google Text-to-Speech. Google's Text-to-Speech consists of two types: Google Text-to-Speech Android Application and Google Cloud Text-to-Speech. The Android application is a feature available in Google's Speech Services app and is free to use. It can assist users by reading text displayed on a smartphone screen aloud. For example, this feature can be used in talkback and accessibility-related apps to provide auditory feedback across user devices.

Previous studies have utilized technologies capable of detecting, reading, and converting text into speech. For instance, a study titled "Comparative Study of Text-to-Speech Engines for Accessibility Applications" [2] focused solely on various TTS engines without exploring direct integration with detection and OCR systems. Another study, "An Evaluation of EasyOCR for Multi-language Text Recognition" [3], evaluated EasyOCR for multiple languages but did not integrate it with detection and TTS systems simultaneously. A third study, "Text Detection and Recognition in Natural Scenes using YOLOv7 and Tesseract OCR" [3][4] concentrated on text detection using YOLOv7 and Tesseract OCR but lacked real-time TTS integration. To address these gaps, this study proposes assistive technology for people with disabilities by combining various applications that complement each other, such as using YOLOv8 to detect images or text. The image or text detection system also employs the EasyOCR artificial intelligence model [4][5]. EasyOCR, an open-source AI model, utilizes CRNN (Convolutional

Recurrent Neural Network) algorithms [6]. It is trained to recognize Latin characters within an image. Based on this capability, EasyOCR is selected as a tool for reading labels in text or images.

Method

The writing begins with a study of relevant literature and then continues with the collection of data that will be used to create the model, in the form of datasets from Roboflow. After that, the author uses OpenCV as an image analysis tool. The authors used models from Ultralytics to train our dataset. Easy OCR helps us in recognizing and extracting text from images. Furthermore, we use GTTS (Google Text-to-Speech) to convert the text generated by Easy OCR into voice speech. The flow below is the whole step-by-step from inputting an image to outputting a voice (Figure 1).

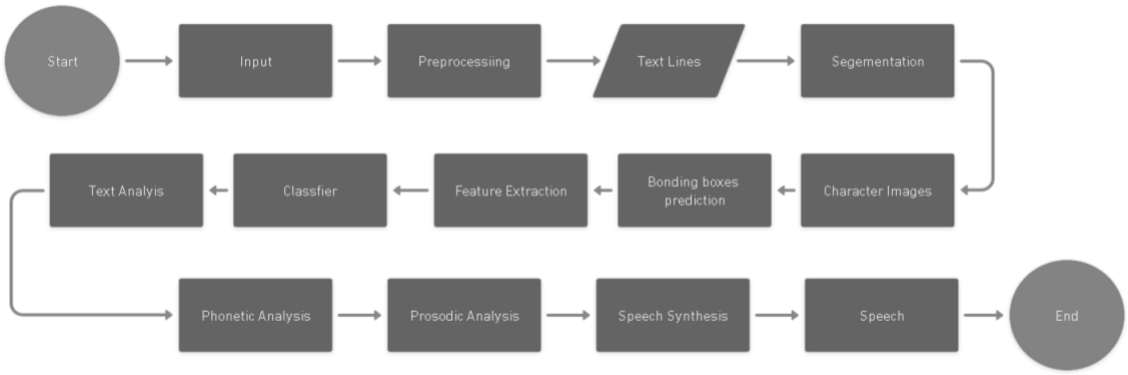

Figure 1. Model of Collection data

Dataset

An image with text dataset is a type of dataset that combines image and text information (such as captions or labels) in a single data set. Each entry in this dataset usually consists of a visual image and text that describes or identifies the image. Image with text datasets is essential in the development of artificial intelligence systems, especially in tasks that require understanding visual and text content simultaneously. It enables the development of models that can relate images to text, helping computers to better understand the relationship between the visual world and human language.

Figure 2. Steps of Dataset

The flow above explains the steps of the Dataset, namely: First take data from RoboFlow, then the data is trained on Google Collab and then the accuracy graph is obtained (seen in Figure 2).

OpenCV

Open Sources Computer Vision (OpenCV) is a powerful software used to perform real-time image processing [7]. OpenCV can be integrated with various programming languages such as C, C++, Java, Python, and Android. In addition, OpenCV provides a collection of patent-free image processing algorithms, which can be used without any legal restrictions preventing their use. OpenCV also enables the detection of lines and circles, as well as features and patterns in images. In addition, the library supports advanced image processing such as image segmentation, feature matching, Fourier transform, and texture analysis. OpenCV's diverse capabilities allow users to effectively detect images in a variety of contexts, from simple tasks to more complex object recognition applications.

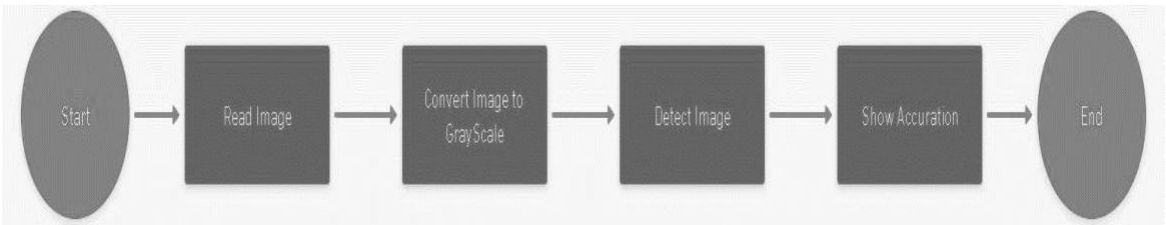

Figure 3. Step of Using OpenCV

The flow above explains the steps of using OpenCV, the explanation is: First read the image, then convert it into GrayScale, then the image is detected and the last one shows the accuracy (seen in Figure 3).

EasyOCR

EasyOCR is an open-source artificial intelligence model, which uses the CRNN (Convolutional Recurrent Neural Network) algorithm [6]. EasyOCR is a model that has been specially trained to recognize Latin letters in images. With this capability, EasyOCR becomes a very useful tool for reading labels or text that appear in images or documents. In various contexts such as text recognition in product photos, document scanning, or character recognition in images, EasyOCR can help efficiently and accurately extract the text present in images, making it a very useful choice in various applications that require text processing in a visual context.

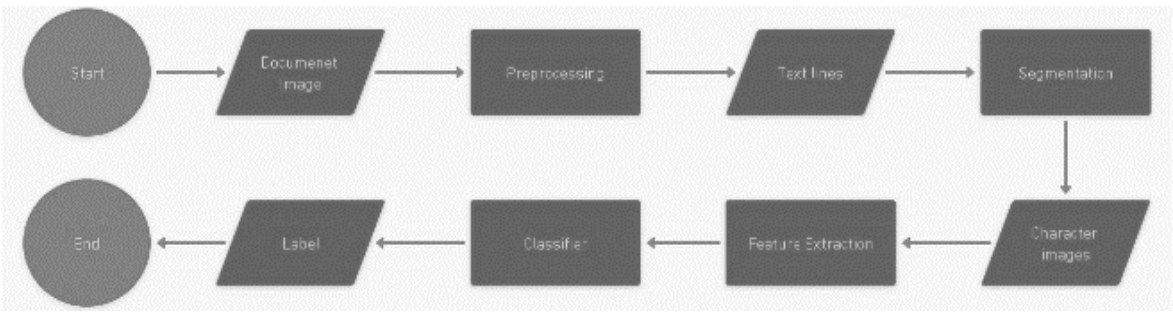

Figure 4. Step of Using EasyOCR

The flow above explains the steps in using EasyOCR. The steps are: the first is the document image, then pre-processing, then the text line and segmenting, then trying on the next character image extracting features, classifying and labeling (Figure 4).

GTTS

Google Text-to-Speech (GTTS) is a screen reader application designed for the Android operating system, and it is powered by Google. Its primary function is to enable applications to audibly read aloud the text content displayed on the screen in multiple languages. Applications like Google Translate utilize GTTS to articulate translations with precise pronunciation and natural intonation, along with numerous other applications that leverage advanced Artificial Intelligence (AI) technologies. Google Cloud Text-to-Speech is generated using WaveNet, a software created by DeepMind, a UK-based AI company acquired by Google in 2014. GTTS offers a diverse selection of voices, encompassing both standard voices and the superior-quality WaveNet voices, which are known for their exceptional quality and natural sound.

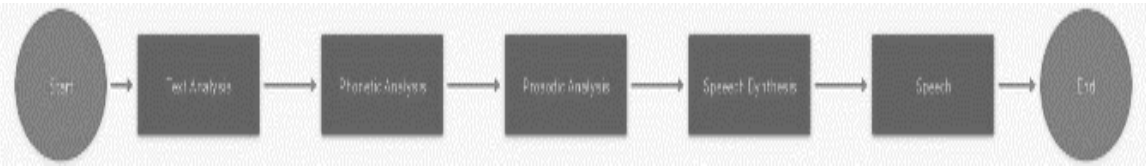

Figure 5. Step Of Using GTTS

The flow above explains the steps in GTTS, the steps are: first analyzing the text, then phonetic and prosodic analysis, then analyzing the speech of the text and finally the speech synthesis (Figure 5).

Result and Discussion

Accuracy results of the training dataset for detecting an image

When training a dataset in Google Collab, several accuracy results will appear. These include training accuracy, which measures the model's ability to understand the training data, validation accuracy, which gives an idea of how well the model can generalize to data it has never seen before, and testing accuracy, which measures the model's performance on completely new data. These accuracy results provide important insights into the quality of the model and help monitor the possibility of overfitting (if training accuracy is higher than validation accuracy) or underfitting (if both accuracies are low). Accuracy evaluation is a key step in the model training process and helps in customizing the model to achieve optimal performance in various scenarios. Accuracy result seen in Figure 6.

All classes in the precision-recall curve with a value of 0.737 in-text detection indicate that the model has consistent precision and recall rates for all text classes analyzed. This value demonstrates the model's ability to identify text, but further evaluation and contextual considerations may be required to understand the overall performance.

These graphs are essential for monitoring and understanding the performance of the model (Figure 7), helping to make informed decisions regarding model enhancements or further actions within a particular project. In addition, data visualization helps in understanding the distribution and structure of the data, enabling the identification of

potential patterns or anomalies that can provide input for further refinement or adjustment of the model and its features.

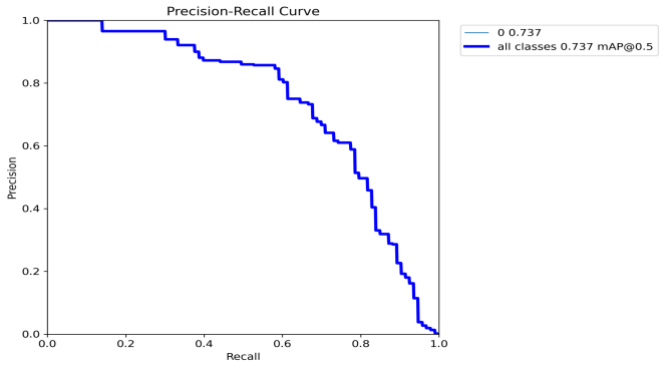

Figure 6. Accuracy Result

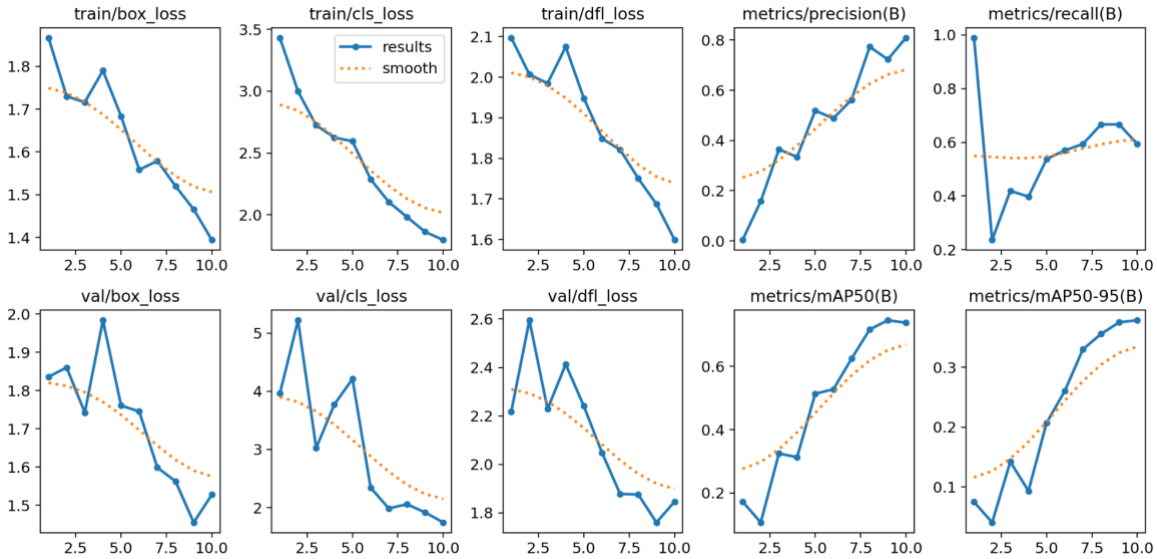

Figure 7. Various Performance

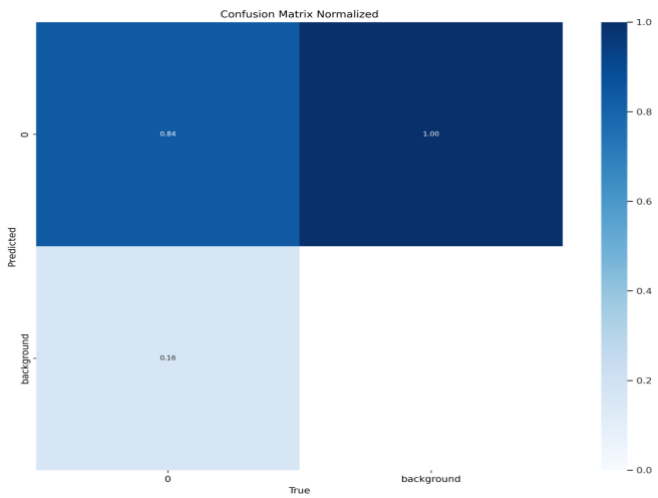

Figure 8. The Normalized confusion matrix value

The normalized confusion matrix values provided offer a clear assessment of the performance of the multi-class classification model (Figure 8). With a perfect accuracy score of 1.00 in the first class, the model excels in classifying all instances in that

category. The second class achieved a good accuracy of 0.84, indicating an accuracy rate of 84%, while the model had difficulty in the third class, with a score of 0.16, reflecting an accuracy of only 16%. These values reveal the strengths and weaknesses of the model in classifying the various categories, prompting deeper examination to understand the factors contributing to the low performance in the third class and to make informed decisions for model improvement or customization as needed.

Implementation of YOLOV8 and EasyOCR models with GTTS

Implementing the YOLOv8 model with GTTS for a visually impaired user can enhance inclusivity and accessibility. By using YOLOv8 to detect objects in their environment and then converting the object detection results into text-to-speech using GTTS, visually impaired users can hear and understand what objects are present around them. Thus, this integration provides additional information and enables them to be more independent in their daily lives, such as identifying objects around them with the assistance of audio.

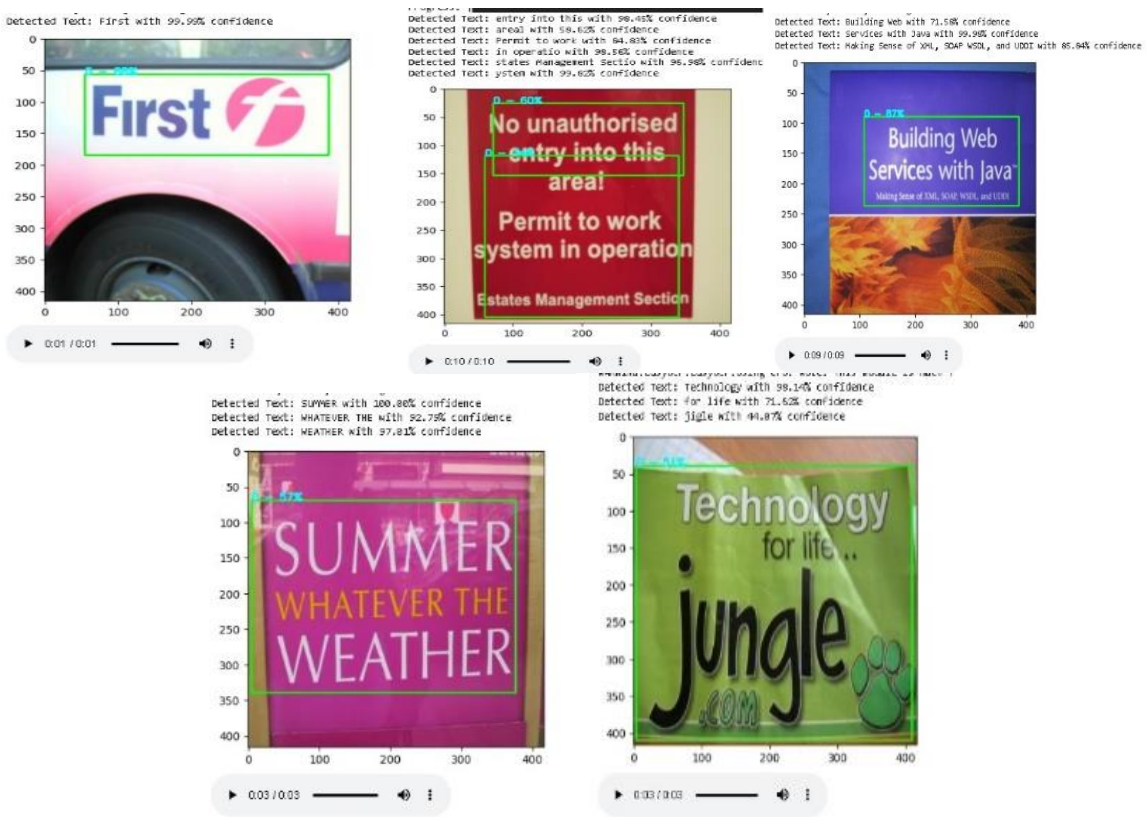

Figure 9. Result of input being entered and producing output in the form of text reading via GTTS

Some of the images above are the result of input being entered and producing output in the form of text reading via GTTS (Figure 9). One example is “Technology for Jungle Life,” where the text is processed individually with accuracy determined by the YOLO model. "Technology" was identified with a confidence level of 98.14%, "for life" with a confidence level of 71.62%, and "jungle" with a confidence level of 44.0%. This allows the text to be spoken clearly and informatively, facilitating a better understanding of the content for the user.

Apart from providing text recognition with high accuracy, the YOLO model also provides confidence for every word detected. This can be useful for users who depend on reading text, such as visually impaired users. With this confidence, users can better understand the extent to which the model believes the words were spoken, which helps them understand the context and information conveyed more fully.

Conclusion

The employment of text detection with GTTS (Google Text-to-Speech) using EasyOCR and YOLOv8 for individuals with visual impairments offers enhanced accessibility for this user group. Through the capabilities of EasyOCR for text detection in images and videos, as well as YOLOv8 for object detection, visually impaired users can listen to the text content present in their environment. With a precision-recall curve accuracy score of 0.737, these results indicate a uniform and consistent model in conveying textual information through voice for visually impaired users. However, for a more comprehensive evaluation, it is necessary to consider other factors, such as class distribution, relative class importance, and specific application context. In the context of equitable access and inclusivity, this solution improves access to text and information for visually impaired users, allowing them to achieve greater independence in various daily activities.

References

[1] Partuni, "Siaran Pers: Peran Strategis Pertuni Dalam Memberdayakan Tunanetra Di Indonesia," *Available:https://pertuni.or.id/siaran-pers-peran-strategis-pertuni-dalam-memberdayakan-tunanetra-di-indonesia/*, Mar. 04, 2017.

[2] A. , & B. H. Smith, "Evaluating Assistive Technology for the Visually Impaired: Methods and Metrics," in *Proceedings of the International Conference on Human-Computer Interaction (HCI)*. , 2019.

[3] X. , & Z. L. Liu, "EasyOCR: A Python Library for OCR with a Focus on Chinese Text.," 2022.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection." [Online]. Available: http://pjreddie.com/yolo/

[5] C. , & Z. H. Wang, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2021, *arXiv preprint*.

[6] Smelyakov, kirill, Chupryna Anastasya, Dmytro Darahan, and Midina Serhii, "Effectiveness of modern text recognition solutions and tools for common data sources.," in *5th International Conference on Computational Linguistics and Intelligent Systems (ICOLINS-2021)*, Ukraina, Apr. 2021.

[7] R. D. , P. W. S. , & T. A. N. Kusumanto, "Aplikasi Sensor Vision untuk Deteksi MultiFace dan Menghitung Jumlah Orang," *Semantik*, 2012.

[8] Jocher, G., & Zhao, D. (2022). YOLOv5: A PyTorch Implementation of YOLOv5. GitHub repository.

[9] Khan, A., & Hussain, M. (2019). A Survey on Optical Character Recognition (OCR) Systems. Journal of Computer and Communications, 7(3), 25-32.

[10] Mishra, S., & Sharma, A. (2018). Text Detection and Recognition in Natural Images Using Deep Learning. In Proceedings of the International Conference on Computer Vision (ICCV).

[11] Rao, K., & Kannan, A. (2020). A Survey of Text-to-Speech Conversion Techniques. Journal of Computer Science and Technology, 35(4), 710-735.

[12] Dhanasekar, D., & Banu, N. (2018). Text to Speech Conversion using Google Text to Speech API. International Journal of Engineering and Technology, 7(2), 823-828.

[13] Arumugam, K., & Ramesh, A. (2021). An Overview of Text-to-Speech Synthesis for Assistive Technologies. In Proceedings of the International Conference on Speech and Language Processing (ICSLP).

[14] Liu, Q., & Wang, W. (2019). Assistive Technology for the Visually Impaired: A Comprehensive Review. Journal of Assistive Technologies, 13(2), 102-119.