HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

Developing a Machine Learning-Based Optical Character Recognition System to Convert Text Images into Audio for the Visually Impaired

Juna Ariyoshi[1] and Jane Chun[#]

[1]Taejon Christian International School,
[#]Advisor

ABSTRACT

Visually impaired individuals face significant challenges when it comes to reading traditional printed books because they rely heavily on visual cues to access written content. Without+ the ability to see, they cannot read the text directly, making conventional reading impossible. As a result, visually impaired people often depend on assistive technologies such as screen readers and audiobooks to access written material. However, converting normal books into audiobooks is a time-consuming, labor-intensive, and expensive process. It involves hiring professional narrators, recording the entire book, and editing the audio to ensure clarity and quality. This process requires significant human and technical resources, driving up costs. To address this problem, I propose a machine learning-based Optical Character Recognition system to convert text images into audio signals. The proposed system utilizes convolutional neural networks and long short-term memory for accurate text recognition and conversion. The approach achieved a word-based exact matching score of 93.724, which is a remarkable result. Furthermore, I implemented the system on a low-cost embedded board to demonstrate its feasibility and applicability in real-world scenarios. I expect that this approach can help visually impaired individuals access written content more easily and affordably.

Introduction

Blindness and vision impairment are emerging as global public health problems. Globally, there are approximately 2.2 billion individuals that bear vision impairment, and among them, 1 billion people suffer with severe visual impairment or blindness (“Vision Impairment and blindness”) (WHO 2023). It is estimated that by 2050, 61 million people will be blind in the United States alone, and the financial expenses from the looming vision loss is expected to reach 373 billion dollars (Bourne et al. 2021). The root cause of this rampant number of blindness lies in the aging population. For instance, South Korea, a country that is officially categorized as an aged society by the United Nation’s population age measures, compels the correlation between increasing number of blindness and aging population. Over the two decades, South Korea’s elderly population nearly tripled; subsequently, the number of blind people recorded during the past two decades shows how from 90,997 individuals, the number surged to 253,055 individuals (Statistics Korea 2020). Visual intake comprises a significant portion of the information we receive about our surroundings and environments. For instance, signs on the street, bus schedules, menus in restaurants, and books for reading all require vision to comprehend and access the information conveyed. The lack of accessibility is a colossal setback that induces the feeling of isolation and disempowerment within the blinds. The blinds not only miss pertinent daily information, they experience difficulty in engaging in leisure activity such as reading.

While there had been efforts to convert books and printed materials barrier-free, these efforts are incompetent. Braille and audio books intended to convert books and printed materials barrier-free for the blind have been incompetent due to the costliness. On average, braille transcription service prices range between $30 to $50 per page, and braille display devices, machines that generate braille real time with texts from USB pairing, range between $3,500

to $15,000 (Blind in Mind 2023) (American Foundation for the Blind 2024). In addition, audio books tend to cost twice as much as paperback copies. The expensive cost of audio and braille books are due to its complex production process that is time-consuming and labor-intensive. Braille transcription and printing requires meticulous braille embossers and specialized, heavier paper. Similarly, producing audio books has a high cost of labor because it requires narrators to record for long hours. The lack of accessibility to books and printed materials seizes the reading rights for the blind where they encounter challenges in securing additional entertainment and educational opportunities.

To address the problem, I introduce a machine learning-based OCR (Optical Character Recognition) system to convert text images into audio for visually impaired patients. The proposed system takes digital text images as input and recognizes text information using OCR networks. The extracted text information is then converted into audio signals via an off-the-shelf text-to-speech module. The rest of this research paper is structured as follows: Chapter 2 explains the modern technology pipeline of OCR and how the system is developed. Chapter 3 provides detailed information on how the proposed system is developed and composed. Chapter 4 presents comprehensive experimental results to prove the feasibility of the proposed methods. Finally, Chapter 5 summarizes the research paper.

OCR

OCR (Optical Character Recognition) is a form of text recognition technology that converts images of typed, handwritten, or printed text into machine-encoded text. The conventional OCR runs based on the pattern matching technique, in which scanned texts are divided into glyphs and are compared with preloaded glyphs in the database. However, as the database is not adaptive, it failed to recognize or misinterpret texts that varied from its set parameters. Furthermore, the OCR heavily relied on the overall quality of the extracted text. For instance, the skewness, color, size, and grade of the text influenced the inaccuracies being made. In response to these limitations, modern OCR utilizes deep-running-based algorithms. OCR runned through deep-running based algorithms is carried out in four different stages.

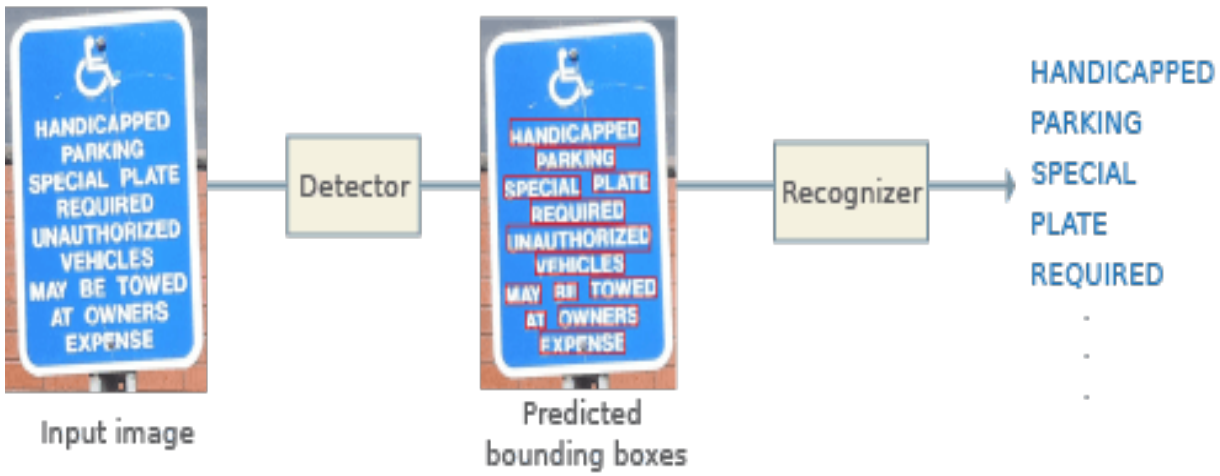
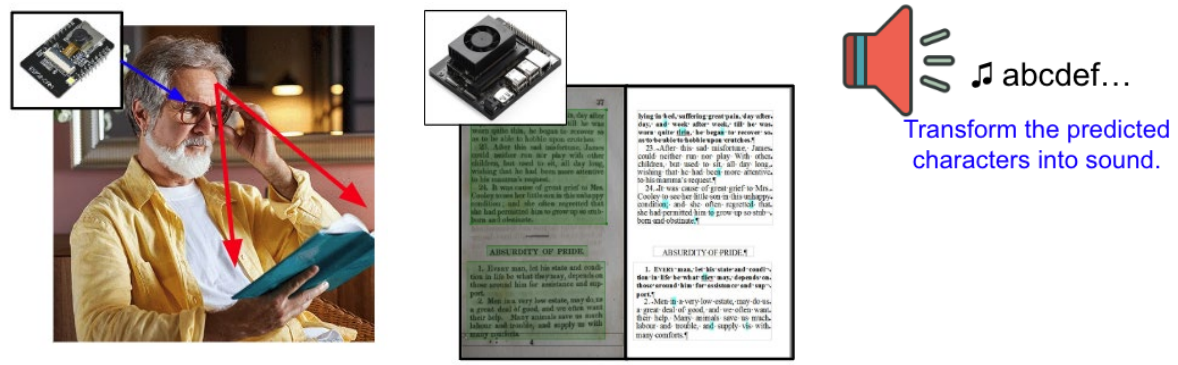


Figure 1. Pipeline of OCR system (MathWorks 2024)

Primarily, the preprocessing stage takes place to adjust the image to enhance recognition accuracy. Computers are unable to distinguish character inherently; rather, they evaluate color differences to analyze pixels. Pixels with similar brightnesses can be perceived as one mass; therefore, the distinct color difference of the regions of texts ensures higher accuracy. The most widely used techniques in the pre-processing stage include converting color images to grayscale, assessing pixel values to amplify brightness and contrast, and executing binary conversion by dividing pixel values into two ranges (0 and 1). The preprocessing stage is followed by text detection and text recognition. The text

detection stage is when the deep-running system identifies the region in which the text is located within the given image. The located text is then divided into each character or/and glyphs to be analyzed in the text recognition stage. With a vast amount of data provided, the computer learns to distinguish various features of the text to be able to recognize it when it comes across it again in the future. Finally, the post-processing stage allows the output text encode to be refined. Through iteration, the deep-running algorithms gain the aptitude to review the logistic of text flow or word, further enhancing the accuracy.

Methodology



(a) (b) (c)

Figure 2. (a): Camera Attached to Eyewear, (b): Jetson embedded board, and (c): Generated Audio Output

Figure 2 illustrates the overall technical flow of the proposed system. As shown in Figure 2 (a), the eyewear-mounted camera collects the visual information which is sent to the Jetson embedded board. Subsequently, as demonstrated in Figure 2 (b), the Jetson embedded board performs OCR where the detected text is converted into machine-readable text format. Finally, through text-to-speech (TTS), the output text encode is converted to audio. (Figure 2 (c)). The following subsections of this chapter are organized as follows: Section 3.1 and 3.2 provide a detailed explanation of the OCR operation, while Section 3.3 offers an in-depth overview of how each module integrates.

Text Detector

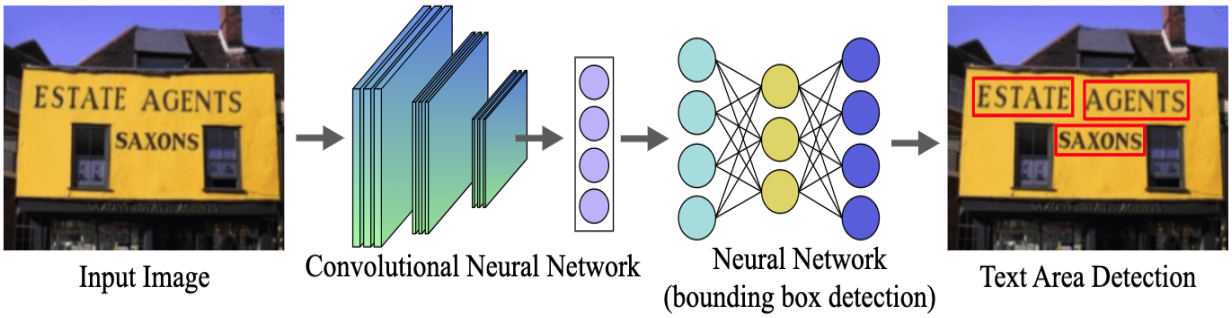


Figure 3. Network architecture of the text detector

Text detector locates segments of texts within an image through forming bounding boxes, as depicted in Figure 3. Through the two-dimensional Gaussian function, the likelihood of a pixel being the center of each character is predicted. The result of the probability map can be visualized through a heat map in which areas of higher and lower likelihood are differentiated. The pixels with the higher probability value are highly probable to be included in the text region.

Text Recognizer

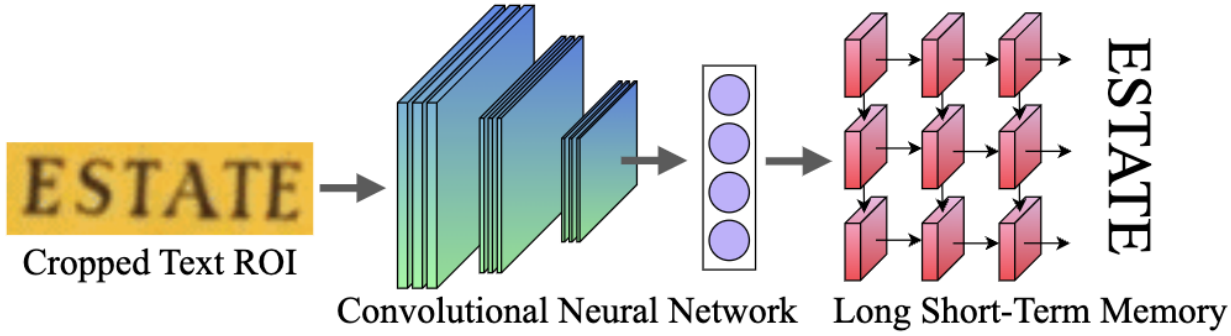


Figure 4. Network architecture of the text recognizer

The text recognizer takes the cropped image and carries out the convolutional neural network (CNN). CNN uses convolutional layers to extract the features from the inputted image. Through the repetition of convolve operation, feature maps are generated. After feature extraction, to model sequences of varying text lengths and order, LSTM (Hochreiter and Schmidhuber 1997) is performed. LSTM is a recurrent neural network that takes the feature maps created by the CNN as the input. As shown in Figure 5, the previous hidden state and the input combine into one matrix and is multiplied with a weight matrix, creating a vector. The vector is then divided into four memory cells (*f, i, g, o*). There are two outputs created; the current cell state is generated through the addition of the product of vector *f* and the previous cell state and the product of vector *i* and *g*. The current hidden state is then generated through the multiplication of vector *o* and the processed value of the current cell state through the hyperbolic tangent function. Then, the current hidden state is multiplied with a weight matrix to determine the output.

Equation 2: Basic operation of LSTM

$$C_t = f \cdot C_{t-1} + i \cdot g$$
$$h_t = o \cdot tanh(C_t)$$

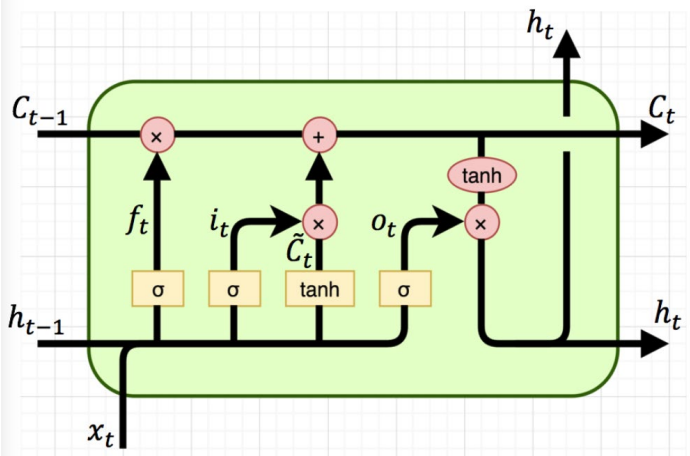


Figure 5. Flow chart of Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber 1997)

Ultimately, in the transcription process, a fixed length probability distribution is created from connectionist temporal classification (CTC). For each position in the sequence, the probability of each character being placed in that position is calculated over a fixed set of characters. The CTC decodes the sequence by utilizing blanks and removing repeated characters to create an accurate textual output.

Experimental Results

OCR Dataset

Three different data sets were utilized for this experiment, with a total of 1,502,089 samples. The first dataset included 400,864 samples in multilingual texts: samples with a mix of Korean and English, a mix of Korean and Chinese, and a mix of Korean and Japanese (AI Hub 2023). The second data set was composed of 50,000 pictures of book covers with some samples having shadows overlapping the texts, adding complexity to the detection process (AI Hub 2023). The final dataset included 1,101,225 samples of texts in both printed and handwritten format.

Figure 6. Samples of dataset used in this paper

Evaluation Metric & Protocol

To assess the performance and the accuracy of the proposed system, I utilized the Word-based Exactly Matching (WEM) evaluation metric where the system looks for an exact match of words between two sets of texts. Any variations are considered incorrect and the system assigns a score of 0 whereas the texts that are identical to the ground truth are awarded the score of 1. The accuracy of the proposed system is then represented through a percentage by dividing the total number of correct matches (scored as 1) by the total number of trials, then multiplying by 100.

Evaluation Results

Table 1. Two groups broken down with age ranges and the difference.

Methodology	**WEM**
Tesseract OCR (Čakić et al. 2020)	88.838
Light OCR (Jenti 2021)	90.709
Easy OCR (Jaided AI 2023)	91.147
Proposed	93.724

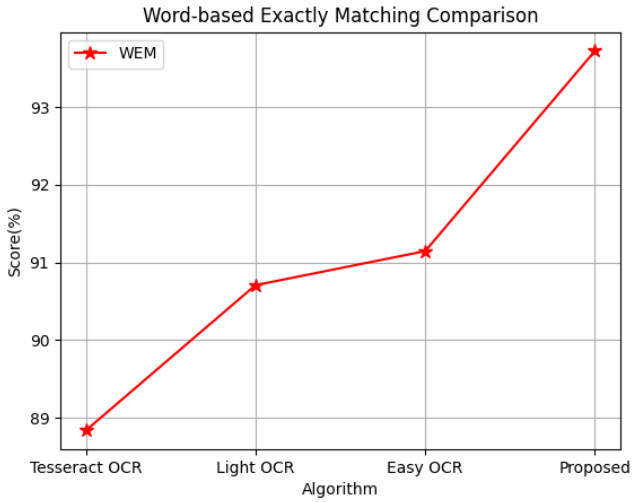


Figure 7. Evaluation result (WEM score)

A series of tests were conducted to evaluate the performance of the proposed method of using the CNN-LSTM based approach. The test was carried out by comparing the performances of the proposed method with other previous state-of-the-art methods, given the same datasets and experimental setup. While the three former approaches that were compared displayed solid performance with relatively high WEM score, the result indicated that the proposed method delivered better outcomes with the highest WEM score of 93.724, a 2.577 score difference from the highest among the three prior methodologies. The results affirm that the proposed method outperforms in terms of performance and accuracy.

Figure 8. Evaluation result (quantitative result)

Conclusion

In this research, I presented a new adaptation for the Optical Character Recognition (OCR) technique, integrating it with the Text-To-Speech framework to increase accessibility of the printed materials for the visually impaired. The new approach was established based on CNN and LSTM and included a variety of datasets for testing to enhance the efficiency and accuracy of text recognition when encountered with texts captured in unfavorable format, like in shadows or/and in fonts with spatial dynamics. The testing outcome revealed that the proposed method demonstrated superior performance in comparison to the previous methodologies, showcasing the highest WEM score of 93.724. Going forward, I seek to elevate the performance of this approach further by exploring additional factors that may influence the outcomes. In the future, I plan to focus on researching practices that can accurately identify texts from images in low resolution.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

AI Hub. (2023, Aug 22). "*Multilingual OCR data*": AI Hub. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71730

AI Hub. (2023, Apr 14). "*Outdoor real shot Korean image*": AI Hub. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=105

American Foundation for the Blind. (2024, Oct 12). "*Refreshable Braille Displays*": American Foundation for the Blind. https://www.afb.org/node/16207/refreshable-braille-displays#:~:text=The%20price%20of%20braille%20displays,the%20number%20of%20characters%20displayed

Blind in Mind. (2024, Oct 12). "*Transcribe Textbooks into Braille*": Blind in Mind. http://www.braillebookstore.com/Braille-Transcription#:~:text=Prices%20for%20Brailling%20a%20textbook,us%20for%20a%20firm%20quote

Bourne, R., Steinmetz, J. D., Flaxman, S., Briant, P. S., Taylor, H. R., Resnikoff, S., ... & Tareque, M. I. (2021). Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. The Lancet global health, 9(2), e130-e143.

Čakić, S., Popović, T., Šandi, S., Krčo, S., & Gazivoda, A. (2020, February). The use of tesseract ocr number recognition for food tracking and tracing. In 2020 24th International Conference on Information Technology (IT) (pp. 1-4). IEEE.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

JaidedAI. (2023, May 25). "*EasyOCR*": JaidedAI.

https://github.com/JaidedAI/EasyOCR

Jenti, (2021, Dec 15). "*Light-OCR-API*": Jenti.
https://github.com/jentiai/Korean-Light-OCR-API

MathWorks. (2024, Sep 24). "*Text Detection and Recognition*":MathWorks.
https://la.mathworks.com/help/vision/text-detection-and-recognition.html

Statistics Korea. (2020, Nov 3). "*The language of the visually impaired, Braille? To celebrate Braille Day, the story of the visually impaired through statistics!*": Statistics Korea.
https://blog.naver.com/hi_nso/222134434777

WHO. (2023, Aug 10). "*Blindness and vision impairment*": WHO
https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment