TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 21, Số 9 (2024): 1623-1636
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 21, No. 9 (2024): 1623-1636
https://doi.org/10.54607/hcmue.js.21.9.4118(2024)

Research Article

BUILDING A DOCUMENT READING ASSISTANT FOR THE VISUALLY IMPAIRED

Thai Thi Kim Yen*, Nguyen Thi Thu Ha, Vo Thi Que Tran,
Huynh Ngo My Vy, Tran Hoang Yen Nhi, Ngo Quoc Viet
Ho Chi Minh City University of Education, Vietnam
**Corresponding author: Thai Thi Kim Yen – Email: 4701104250@student.hcmue.edu.vn*
Received: January 28, 2024; Revised: June 13, 2024; Accepted: June 26, 2024

ABSTRACT

This study introduces a solution that applies document analysis and recognition technologies to enhance document accessibility for individuals with visual impairments. The objective is to develop an algorithm capable of accurately analysing the content of document components and converting them into voice format. Leveraging the pre-trained YOLOv8 model for document analysis and optical character recognition technology, the image annotation model uses the AIAnytime API and Pix2Tex technology to extract LaTeX code from images, facilitating the conversion of mathematical formulas into spoken words. The research results demonstrate significant progress in effectively supporting document reading, making a meaningful contribution to assistive technology for the visually impaired.

Keywords: document analysis and recognition; document image processing; visual impairments

1. Introduction

The term "visually impaired" describes a condition that cannot be corrected by medication or surgery. It includes individuals with partially impaired vision as well as those who are completely blind. Visually impaired individuals face numerous challenges in acquiring information visually, particularly when reading books. Initially, a visually impaired person desiring to read a book has no alternative but to rely on someone else to read it aloud. With the progress of humanity, numerous studies have been conducted to normalise the act of reading for individuals with visual impairments. As a result, various innovations and solutions have emerged, including braille books and audiobooks.

The Braille system provides an effective solution for visually impaired individuals when reading books or documents. However, it faces limitations such as slow reading speed

Cite this article as: Thai Thi Kim Yen, Nguyen Thi Thu Ha, Vo Thi Que Tran, Huynh Ngo My Vy, Tran Hoang Yen Nhi, & Ngo Quoc Viet (2024). Building a document reading assistant for the visually impaired. *Ho Chi Minh City University of Education Journal of Science, 21*(9), 1623-1636.

and difficulties in language transitions. Braille materials are limited, and publishing them comes with high costs. Audiobooks can address issues like reading speed without requiring the listener to learn Braille. However, they are often released after printed books and are rarely available for other documents.

An increasing number of research works have been focusing on addressing the challenges faced by the visually impaired. AI-based applications have emerged as promising solutions in specific domains of visual impairment assistance. Wang et al. (2023) examine the use of artificial intelligence in eye disease diagnosis, showcasing the potential for AI to enhance early detection and management of ocular conditions. Additionally, their research also addresses the realm of intelligent assistive devices, highlighting how AI algorithms can be integrated into assistive technologies to improve functionality and usability for visually impaired individuals.

This study focuses on machine learning models and artificial intelligence, along with integrated technological devices, which have been employed to develop applications that support document reading, recognise surrounding objects, and even provide navigation guidance through voice and touch.

Ganesan et al. (2022) use deep learning algorithms like CNN and LSTM to extract features, create image captions, and convert text to speech. Khan et al. (2020) developed a visual assistance system for the blind using Raspberry Pi, featuring a camera, sensor for obstacle avoidance, object detection algorithms, and a reading assistant to convert images to text with auditory feedback. Kapgate et al. (2023) created a Raspberry Pi-based OCR tool to convert text to an audio in real time, recognising various text types.

In researching methods and developing systems to support document reading for the visually impaired, extracting content from images of textual documents plays a crucial role. Faced with diverse formats, layouts, and structures, automating the analysis and segmentation of images requires complex techniques to ensure accurate identification and extraction of essential components. Document Analysis and Recognition (DAR) is a research field within image processing and artificial intelligence, focusing on extracting information from textual documents such as books, articles, and printed material. A crucial task in DAR is document image segmentation, concentrating on separating and identifying components like text, images, formulas, tables, and geometric shapes from the document images. Efforts have been made to standardise ground truth data formats, facilitating training procedures for Object Detection (OD) methods on datasets. Notable large-sized datasets, such as TableBank (Li et al., 2020), DocBank (Li et al., 2020), DeepFigures (Siegel et al., 2018), PubTabNet (Zhong et al., 2020), and PubLayNet (Zhong et al., 2019), have been introduced to support document classification, analysis, and understanding. Additionally, recent datasets like NCERT5k-IITRPR (Kawoosa et al., 2022) focus on text/non-text component analysis, the Laser-Printed Characters Dataset (Furukawa, & Takeshi, 2021)

addresses document forensics, and DocLayNet (Pfitzmann et al., 2022) serves for general-purpose Document Layout Analysis (DLA).

To assist visually impaired individuals in overcoming difficulties with document reading, Wang et al. (2021) introduced SciA11y, which integrates multiple machine learning models to extract content from scientific PDFs and convert it into accessible HTML. The study by Fayyaz et al. (2023) proposes a method to extract explicit and implicit features, including metadata, functional, structural, content, and contextual information, by providing in-depth insights into PDF tables.

Among the related studies, significant progress was observed in prior works, primarily focusing on text processing while paying less attention to handling image components, formulas, and other non-textual elements. This motivation has prompted addressing and enhancing this aspect in developing the solution for this study.

This paper introduces a solution to empower visually impaired individuals by facilitating document reading through advanced Document Analysis and Recognition (DAR) techniques. Images of the document pages will be captured, and the content components will be analyzed based on DAR. Each component will be processed using different methods, to convert them into spoken format.

This research aims to significantly assist visually impaired individuals in becoming more independent in reading books and conducting research. It not only helps them avoid dependence on assistance from others to access the content of books but also has the potential to address the limitations of Braille systems and audiobooks. This promises to open up new opportunities for them to access information and participate in academic and research activities.

2. Research Content

2.1. Proposed Method

To facilitate efficient access to reading materials for individuals with visual impairments, the conversion of document images into an auditory format is undertaken. In this approach, the document image segmentation method is employed to process each document component individually. Beyond the extraction and processing of text, emphasis is placed on non-textual elements such as images and mathematical formulas. An image annotation process is undertaken for each non-textual component, and mathematical formulas are converted into a comprehensible form, presenting them in corresponding text. Ultimately, all these components are integrated and pronounced, aiming to optimise the document reading experience for individuals with visual impairments. Figure 1 illustrates an overview of this approach.

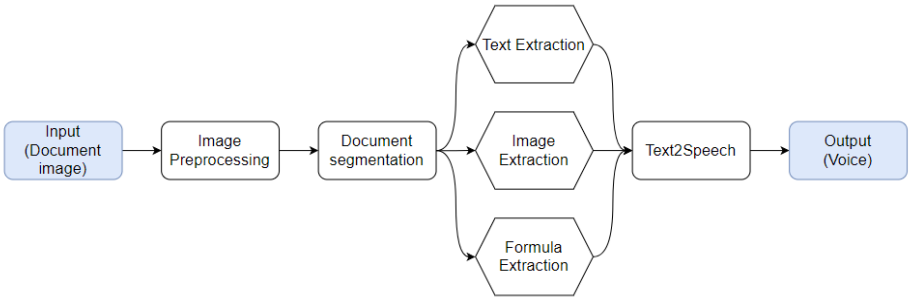

Figure 1. *The process of converting document images into speech*

The input data comprises images containing pages of documents along with other components in the background, posing a challenge in document recognition. To address this challenge, a method is proposed to accomplish this task.

This method uses morphological operations and the K-means algorithm to classify objects and separate them from the background (Figure 2). The preprocessing step of morphological close operation removes small, insignificant details, retaining essential features, thereby enhancing input data quality and document recognition.

Figure 2. *The image preprocessing steps are based on the K-means clustering algorithm*

In the next step of this research, the document images are segmented after the preprocessing stage. This process aims to divide the image into different regions, each corresponding to one of the five classes: *text, images, formulas, tables, and geometrical shapes*. This helps to define the positions of the elements within the image clearly and provides information about their relationships, making extracting content from the document components more efficient.

Using the pre-trained YOLOv8 model has proven to be an effective and versatile strategy. Initiated by training the model on a large and diverse dataset, it not only enables the model to learn common features of objects but also enhances its generalisation capabilities. Once these features are learned, the model undergoes fine-tuning on specific datasets, adapting well to the specific characteristics of the object recognition task within documents. This adaptive process boosts performance and reduces the demand for extensive training data, saving time and effort in the development process.

Within the scope of this study, the extraction of content from segmented text, images, and formulas is performed.

To efficiently carry out the process of extracting text from images, OCR technology is integrated into the workflow. The OCR utilisation begins with meticulous image preprocessing steps. Initially, the image is converted to grayscale, optimising the input for subsequent OCR processing. Following this, sophisticated thresholding techniques are applied to enhance the contrast and improve the OCR model's ability to discern text accurately. Subsequently, an OCR model is deployed to effectively extract textual content from the preprocessed image. This ensures that the process transforms images into readable text and guarantees high accuracy and performance in extracting information from images. Figure 3 describes the process of extracting text from images.

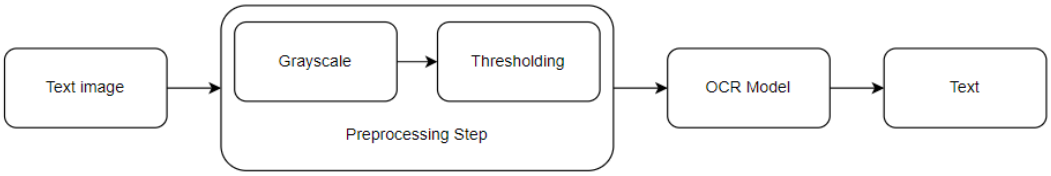

Figure 3. *Text extraction process*

Images are often non-text elements, so OCR technology cannot effectively process them. In the case of images, an image captioning model is proposed to generate textual descriptions, facilitating the expression and communication of essential features of the image in a language that can be easily pronounced. In image captioning, the image colour space is converted from the original colour space to the RGB colour space. This transformation is crucial, shifting the image's original colour to the widely recognised RGB colour space. Subsequently, an image captioning model is integrated based on an encoder-decoder architecture. Figure 4 depicts the process of handling image components.

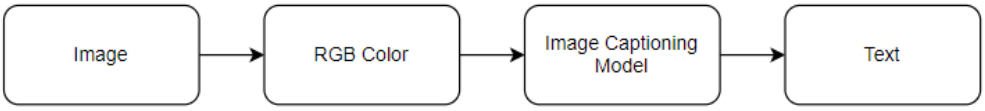

Figure 4. *Image extraction process*

Extracting mathematical formulas and converting them into readable text for comprehension is entirely different from using OCR. Mathematical formulas often contain complex elements and specific relationships between characters and expressions. Directly applying OCR to formulas cannot ensure an accurate understanding of their structure and meaning.

A formula recognition model is utilized to process mathematical formulas from images, and the output is represented in LaTeX formula format. However, to ensure that these formulas can be understood and read conveniently, a crucial intermediate step is undertaken, which is the conversion from LaTeX format to MathML (Mathematical Markup Language) format. MathML helps describe the syntax and content of mathematical expressions in a detailed and understandable manner for computers and other applications.

Next, a speech rules tool is used to analyse the components in MathML format and convert them into text that reflects the meaningful relationships between the elements in the formula. This ensures that information about the structure and meaning of the formulas is preserved and conveyed accurately. The entire process of extracting and processing the formula is described in Figure 5.

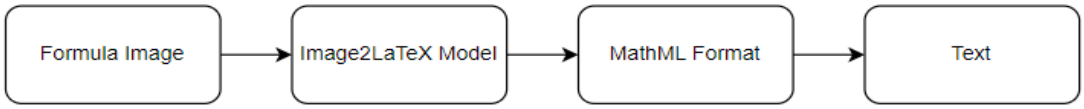

Figure 5. *Formula extraction process*

Converting document components into audio format is an optimal and convenient solution to facilitate efficient access to reading materials for individuals with visual impairments. After completing the processing steps for the document components and generating a comprehensive text, the conversion process proceeds, transforming the text into audio.

2.2. Experiments

2.2.1. Dataset

Textbooks are the primary data source due to their accuracy, fundamental nature, and organizational structure. They serve as official and mandatory learning materials for students and educators within the education system.

This dataset comprises images scanned from the content of Grade 9 Mathematics textbooks Volume 1 by Phan Đức Chính, Tôn Thân, Vũ Hữu Bình, Trần Phương Dung, Ngô Hữu Dũng, Lê Văn Hồng, and Nguyễn Hữu Thảo; Volume 2 by Phan Đức Chính, Tôn Thân, Nguyễn Huy Đoan, Phạm Gia Đức, Trương Công Thành, and Nguyễn Duy Thuận; Grade 9 Mathematics Exercise Book Volume 1 by Tôn Thân, Vũ Hữu Bình, Trần Phương Dung, Lê Văn Hồng, Nguyễn Hữu Thảo; Grade 9 Mathematics Exercise Book Volume 2 by Tôn Thân, Phạm Gia Đức, Trương Công Thành, and Nguyễn Duy Thuận; and Grade 9 Biology textbook by Nguyễn Quang Vinh, Vũ Đức Lưu, Nguyễn Minh Công, and Mai Sỹ Tuấn. These books were published by the Education Publishing House in 2005.

The fundamental dataset consists of 895 images, each representing a single page from the textbooks. The data is categorised into five classes, each representing a specific type of content on the textbook pages. The classes include:

1. Formula: Content containing mathematical expressions, equations, or scientific formulas.
2. Geometry: Content featuring geometric images, such as spatial geometry, plane geometry, and related illustrations.
3. Image: Content comprising illustrative images, primarily visuals related to biology content or real-world examples in Mathematics textbooks.
4. Table: Content containing data tables and figures.

5. Text: Content consisting of pure text.

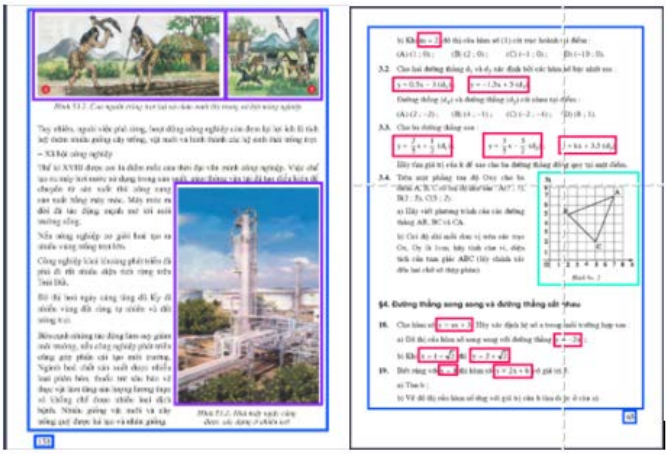

Figure 6. *Textbook page images have been labeled*

Figure 6 illustrates the labelled images—applied preprocessing techniques, including Auto-Orient and Resize (stretched to 640x640), to optimise the dataset. Additionally, augmentations were implemented to enhance the diversity and robustness of the dataset. These augmentations encompassed various transformations such as Crop, Rotation, Shear, Grayscale, Cutout, Mosaic, Brightness adjustment, and Blur. The chart displaying the count of labels for each class is presented in Figure 7.

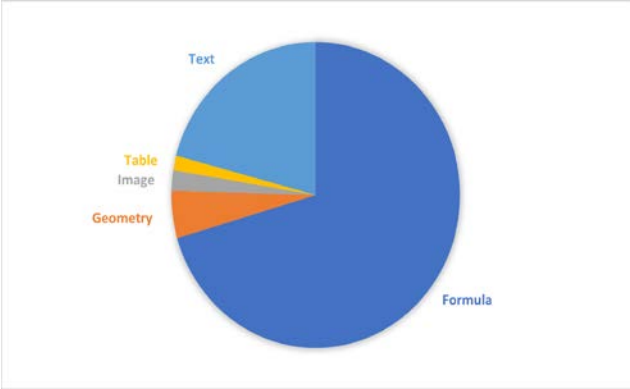

Figure 7. *The number of labels for each class in the dataset*

Finally, a dataset comprising 1300 images has been constructed, with 1032 images in the training set, 179 in the validation set, and 89 in the test set. Samples of the dataset are shown in Figure 8.

Figure 8. *Some samples in the dataset*

2.2. Experiment Design

A pre-trained YOLO model has been employed for segmentation purposes. This method involves fine-tuning the YOLOv8 model on the specific dataset, which includes images with annotated segmentation masks. The YOLOv8s-seg model is utilised, comprising 261 layers with 11,792,031 parameters and 11,792,015 gradients. The SGD optimisation algorithm is employed during the training process with a learning rate of 0.01. The image size used for both training and evaluation is 640x640 pixels. Two workers are activated when using DataLoader to optimise data loading speed. The training process is expected to run for 100 epochs.

The OCR model utilised in this approach is Tesseract OCR, one of today's most powerful and widely used tools. This OCR model was chosen for its excellent character recognition capabilities, flexibility, and customisation options, allowing adaptation to various formats and types of documents. The stability and accuracy of Tesseract OCR have significantly contributed to extracting text from images, ensuring the precision and completeness of the obtained information. The language experimented on in this project is Vietnamese.

In the image captioning model, an API developed by AIAnytime is utilised, optimised explicitly for image caption generation. This API integrates FastAPI and the Vision Transformer (ViT) model. The decision to integrate this API into the approach is based on its advanced capabilities in generating descriptive captions for images.

In extracting formulas from images, the Pix2Tex library and the LatexOCR model are utilised to recognise LaTeX code. Subsequently, the LaTeX code undergoes conversion into the MathML format. Then, using the Speech Rule Engine (SRE) through JavaScript, MathML is transformed into speech, generating detailed descriptions of the mathematical formula content.

We are consolidating and rearranging the results of OCR, image annotation, and formula annotation processes to generate a cohesive text. This comprehensive text is then transformed into audio using Google's Text-to-Speech API, providing a convenient document reading experience for individuals with visual impairments.

3. Results

Figure 9 visually represents the outcomes observed at various stages during the model training process. Additionally, Table 1 presents a comprehensive overview of the average loss values throughout training.

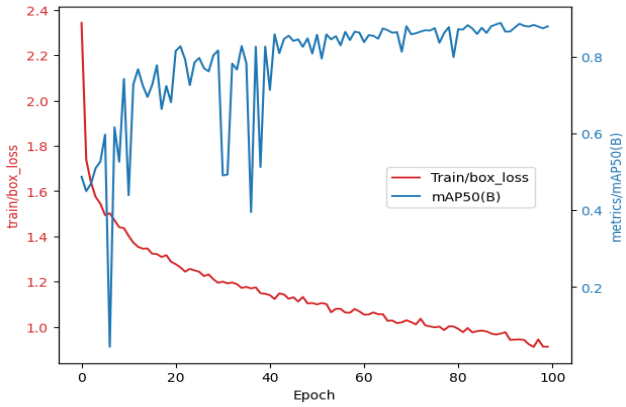

Figure 9. *Results during the model training process*

The average loss values during the training process show that the model achieves excellent classification performance with an Avg Class-Loss of 0.912, indicating that the model's object recognition capability is very accurate. Although there are still some challenges in segmentation with an Avg Seg-Loss of 1.561, predicting positions and sizes with an Avg Box-Loss of 1.154, as well as specific points with an Avg Dfl-Loss of 1.122, these values are all at an acceptable level.

Table 1. *The average loss values during the training process*

Avg Box-Loss	Avg Seg-Loss	Avg Class-Loss	Avg Dfl-Loss
1.154	1.561	0.912	1.122

Table 2 represents the performance evaluation metrics of the model, such as mAP, Precision, and Recall for the recognition and localisation of bounding boxes across different classes. The "Table" class achieves the highest mAP50 with a value of 98.3%, which is an exceptionally high metric. With mAP50 at 77.9%, Precision at 82.8%, and Recall at 75.7%. However, the mAP50 is reasonably good; the difference between Precision and Recall may highlight the challenge for the model in maintaining a balance between accuracy and evaluation for the "Text" class.

Table 2. *Model performance evaluation in recognising and localising bounding boxes*

	Formula	Geometry	Image	Table	Text	All
mAP50	88,2%	96,4%	80,5%	**98,3%**	77,9%	88,3%
Precision	83,4%	91,7%	72.6%	96,4%	82,8%	85,4%
Recall	85%	91,4%	82,9%	90,2%	75,7%	85%

Figure 10 depicts the confusion matrix among the classes. Overall, the model performs well across multiple classes, but some areas need improvement, particularly for the "Image" and "Text" classes. The model performs excellently for the "Table" and "Geometry" classes. The model may require fine-tuning and optimization for other classes to enhance accuracy and evaluation.

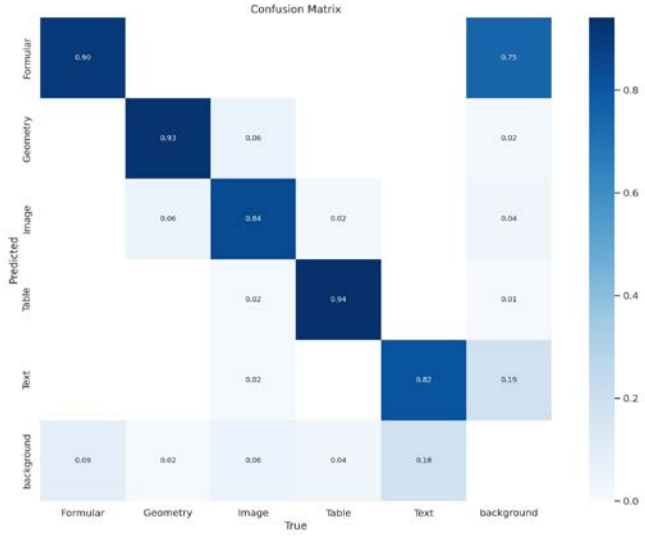

Figure 10. *Confusion matrix*

Table 3 compares results with related works, including different models, datasets, and evaluation metrics such as mAP@IoU0.6, mAP@IoU[0.50:0.95], and mAP@IoU0.5. The respective values indicate the performance of each model on the specified datasets.

Table 3. *Compare the results with related works*

Model	Dataset	Evaluation Metrics	Value
NLPR-PAL (Gao et al., 2017)	ICDAR-POD	mAP@IoU0.6	87,4
Mask R-CNN (Zhong et al., 2019)	PubLayNet	mAP@IoU[0.50:0.95]	91
Pre-train YOLOv8	Textbook	mAP@IoU0.5	88,3

Figure 11 illustrates document images before and after the segmentation process. In this image, the outcome of document segmentation is observed, revealing distinct components and structures within the document after processing.

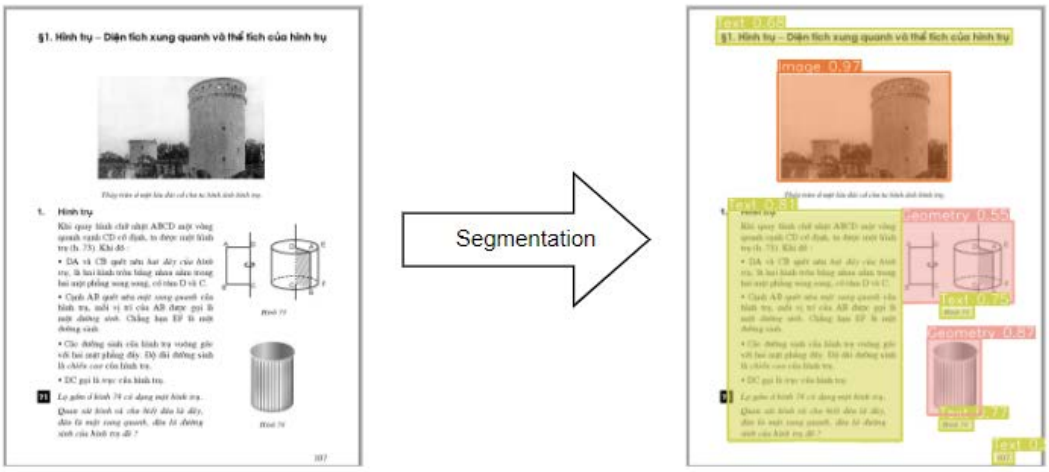

Figure 11. *Document images before and after segmentation*

Figure 12 presents the result obtained from the specifically optimised OCR process for handling Vietnamese text. This outcome showcases the efficiency of the automatic handwriting recognition technology using the Tesseract OCR tool, demonstrating the system's effective processing and decoding capabilities for Vietnamese text.

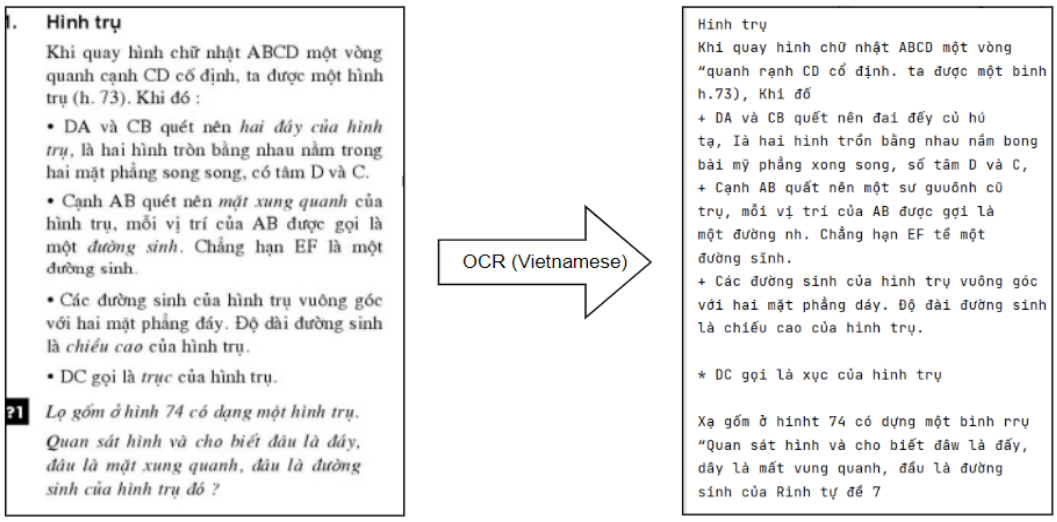

Figure 12. *Results of Optical Character Recognition in Vietnamese*

Figure 13 represents the result of the image captioning process in this solution. Throughout this process, an API developed by AIAnytime is integrated. To meet the specific requirements of the task, the resulting captions are then translated from English to Vietnamese.

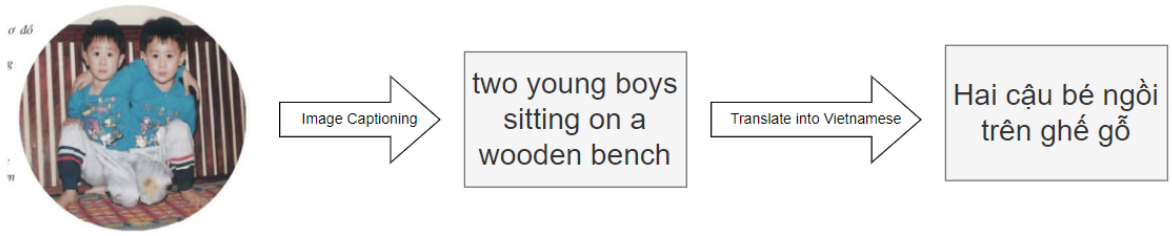

Figure 13. *Results of the image caption generation process*

Figure 14 illustrates the results of generating captions for formulas using LatexOCR technology to recognise formulas and the Speech Rule Engine (SRE) to convert them into speech, providing detailed descriptions of their content.

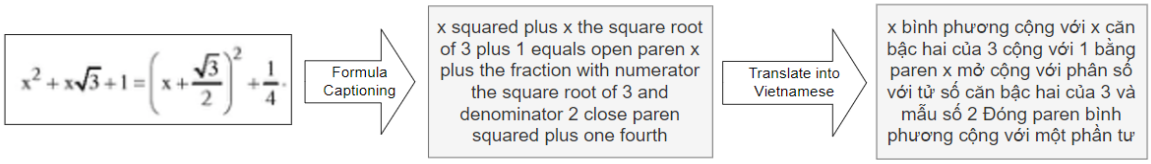

Figure 14. *Results of the formula caption generation process*

The final results are depicted in Figure 15. Consolidating and reorganising the outcomes of OCR, image annotation, and formula processing processes to generate a cohesive text. Ultimately, it is converted into audio.

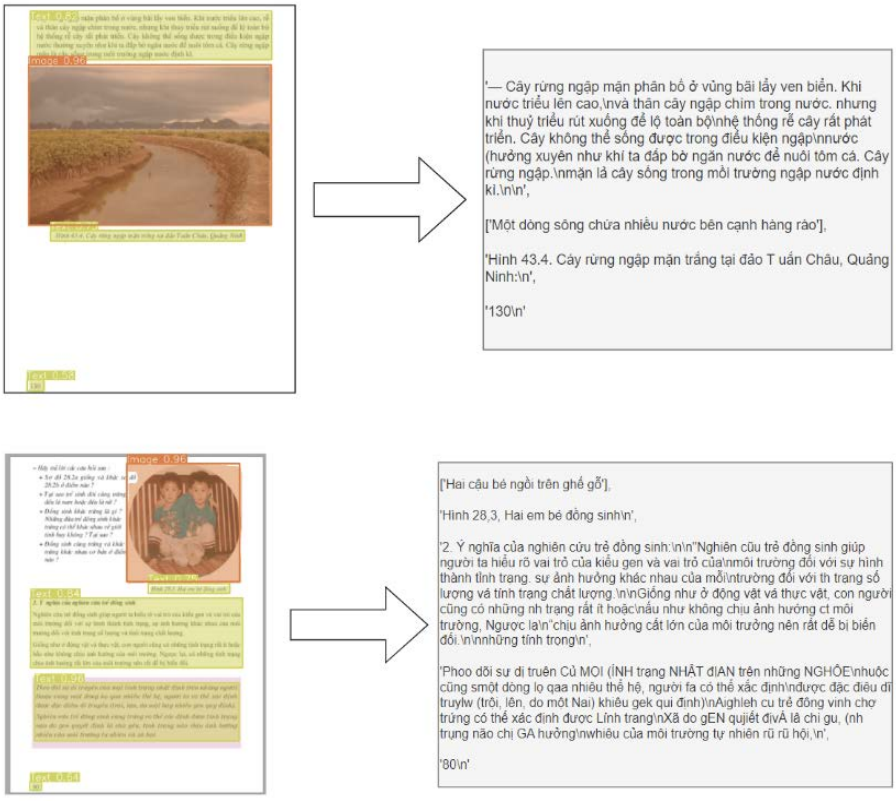

Figure 15. *Final result*

4. Conclusions and recommendations

This method demonstrates high efficiency, mainly when there is limited overlap between components during the segmentation process. Simultaneously, image quality plays a crucial role in ensuring favourable processing outcomes. The subsequent steps in this research focus on enhancing model accuracy and execution time.

In future works, efforts to optimise performance will be made, especially in cases with substantial overlap between various document components. The objectives of this study include processing and annotating remaining components, such as tables and graphics, further advancing the capabilities of this approach. Additionally, this algorithm will be deployed on embedded systems to enhance the practical applicability of this research work.

❖ ***Conflict of Interest:*** *Authors have no conflict of interest to declare.*

REFERENCES

Fayyaz, N., & Khusro, S. (2023). Enhancing Accessibility for the Blind and Visually Impaired: Presenting Semantic Information in PDF Tables. *Journal of King Saud University-Computer and Information Sciences*, *35*(7), 101617.

Furukawa, T. (2021). Recognition of Laser-Printed Characters Based on Creation of New Laser-Printed Characters Datasets. In *International Conference on Document Analysis and Recognition* (pp. 407-421). Springer International Publishing.

Ganesan, J., Azar, A. T., Alsenan, S., Kamal, N. A., Qureshi, B., & Hassanien, A. E. (2022). Deep learning reader for visually impaired. *Electronics*, *11*(20), 3335.

Gao, L., Yi, X., Jiang, Z., Hao, L., & Tang, Z. (2017, November). ICDAR2017 competition on page object detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 1417-1422). IEEE.

Kapgate, P., Tidke, S., Fender, R., Rathore, S., Ghodmare, S., & Ritla, L. (2023, April). Raspberry Pi Based Book Reader For Visual Impaired People. In *2023 11th International Conference on Emerging Trends in Engineering & Technology-Signal and Information Processing (ICETET-SIP)* (pp. 1-6). IEEE.

Kawoosa, H. S., Singh, M., Joshi, M. M., & Goyal, P. (2022, May). NCERT5K-IITRPR: A Benchmark Dataset for Non-textual Component Detection in School Books. In *International Workshop on Document Analysis Systems* (pp. 461-475). Springer International Publishing.

Khan, M. A., Paul, P., Rashid, M., Hossain, M., & Ahad, M. A. R. (2020). An AI-based visual aid with integrated reading assistant for the completely blind. *IEEE Transactions on Human-Machine Systems*, *50*(6), 507-517.

Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., & Li, Z. (2020, May). Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1918-1925).

Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., & Zhou, M. (2020). DocBank: A benchmark dataset for document layout analysis. *arXiv preprint* arXiv:2006.01038

Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A. S., & Staar, P. (2022, August). Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3743-3751).

Siegel, N., Lourie, N., Power, R., & Ammar, W. (2018, May). Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 223-232).

Wang, J., Wang, S., & Zhang, Y. (2023). Artificial intelligence for visually impaired. *Displays, 77*, Article 102391.

Wang, L. L., Cachola, I., Bragg, J., Cheng, E. Y. Y., Haupt, C., Latzke, M., Kuehl, B., Zuylen, M. V., Wagner, L. & Weld, D. S. (2021). Improving the accessibility of scientific documents: Current state, user needs, and a system solution to enhance scientific PDF accessibility for blind and low vision users. *arXiv preprint* arXiv:2105.00076

Zhong, X., ShafieiBavani, E., & Jimeno Yepes, A. (2020, August). Image-based table recognition: data, model, and evaluation. In *European Conference on Computer Vision* (pp. 564-580). Springer International Publishing.

Zhong, X., Tang, J., & Yepes, A. J. (2019, September). Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1015-1022). IEEE.

XÂY DỰNG TRỢ LÍ ĐỌC TÀI LIỆU CHO NGƯỜI KHIẾM THỊ

Thái Thị Kim Yến*, Nguyễn Thị Thu Hà, Võ Thị Quế Trân,
Huỳnh Ngô Mỹ Vy, Trần Hoàng Yến Nhi, Ngô Quốc Việt
Trường Đại học Sư phạm Thành phố Hồ Chí Minh, Việt Nam
**Tác giả liên hệ: Thái Thị Kim Yến – Email: 4701104250@student.hcmue.edu.vn*
Ngày nhận bài: 28-01-2024; ngày nhận bài sửa: 13-6-2024; ngày duyệt đăng: 26-6-2024

TÓM TẮT

Nghiên cứu này giới thiệu một giải pháp áp dụng công nghệ phân tích và nhận diện tài liệu để cải thiện khả năng tiếp cận tài liệu cho những người khiếm thị. Mục tiêu là phát triển một thuật toán có khả năng phân tích nội dung của các thành phần văn bản một cách chính xác và chuyển đổi chúng thành định dạng giọng nói. Sử dụng mô hình YOLOv8 đã được đào tạo trước để phân tích tài liệu và công nghệ nhận diện kí tự quang học, mô hình chú thích hình ảnh sử dụng API AIAnytime và công nghệ Pix2Tex để trích xuất mã LaTeX từ hình ảnh, hỗ trợ chuyển đổi các công thức toán học thành từ ngữ nói. Kết quả nghiên cứu chứng minh tiến triển đáng kể trong việc hỗ trợ hiệu quả việc đọc tài liệu, đóng góp ý nghĩa cho lĩnh vực công nghệ hỗ trợ cho người khiếm thị.

Từ khóa: phân tích và nhận dạng tài liệu; xử lí ảnh tài liệu; người khiếm thị