Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 2, July 2025, pp. 538 – 543
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Robustness Testing of TrOCR for Multi-Condition Food Ingredient Labels Detected By YOLO

[1]Charina Mutiara Chairunnisa, [2*]Nyayu Latifah Husni, [3]RD. Kusumanto
[1,2,3]Department of Electrical Engineering, Politeknik Negeri Sriwijaya, Indonesia
Email: [1]charina@polsri.ac.id, [2]nyayu_latifah@polsri.ac.id, [3]kusumanto@polsri.ac.id

Article Info
Article history:
Received May 28th, 2025
Revised Jul 16th, 2025
Accepted Jul 30th, 2025

Keyword:
Ingredient Label
Optical Character Recognition
Text Extraction
Transformer
YOLOv8

ABSTRACT
This study aimed to develop an automatic text extraction system for ingredient labels by integrating YOLOv8 for object detection and a Transformer-based Optical Character Recognition (OCR) for text recognition. YOLOv8 was trained to detect and crop the label area in the image, while TrOCR was used to extract text from the cropped bounding box. The evaluation involved 16 sample image inputs under various conditions, including background color (Monochrome and RGB), languages (Bahasa Indonesia and English), and text formatting (single-line and multi-line). The results indicated that TrOCR performed well in single-line format, but struggled with multi-line format and longer text, even omitting words. Character and word error rates reached up to 100% for this complex layout.

Corresponding Author:
Nyayu Latifah Husni,
Department of Electrical Engineering,
Politeknik Negeri Sriwijaya,
Srijaya Negara St, Bukit Lama, Ilir Barat I District, Palembang City 30128, South Sumatera, Indonesia.
Email: nyayu_latifah@polsri.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i2.37301

1. INTRODUCTION

OCR or Optical Character Recognition is a technology known for its ability to extract text from images. It is inherently categorized as Extractive Artificial Intelligence (AI) due to its performance on the detection and extraction of text accurately. It works as light intensity is converted into electrical signals, and then processed to be machine-readable text [1]. OCR is known for its ability to recognize both handwritten and printed text [2].

In practical applications, OCR plays a pivotal role across various sectors, such as in healthcare for digitalizing medical reports and prescriptions, finance for processing insurance receipts, retail and food industry for price and product labels, paper documents, etc cetera [3]. Thus, OCR is a transformative technology that enhances productivity in automation and detection, eliminating the need for manual entry, as seen in these sectors.

One advanced form of OCR is Transformer-based OCR or to abbreviate TrOCR. This approach leverages a transformer architecture with a self-attention mechanism [4], that has encoders and decoders to break down each character in the text into visual patches and generate text tokens using features from encoders. TrOCR offers various model sizes (small, base, and large models) that are adaptable for different OCR complexities, including handwritten and printed text [5].

As utilized in the food industry, OCR can certainly be relied on to recognize essential product information, such as nutrition facts, allergens, label ingredients, expiration dates, barcodes, and serial numbers in food products. However, many real-world food products have multilingual labels and complex layouts, which can reduce OCR accuracy and make it difficult for consumers to make a quick decision at the point of purchase. That is vital for quality assurance and ensuring consumer safety. By accurately detecting and extracting, it provides good consumer protection and regulatory audits.

Recent studies, Large Model TrOCR [6], [7], have demonstrated decent performance in reading complex 2d engineering and scanned Arabic documents by achieving the lowest error rates. Eventually, it had

limitations regarding special characters, numerals, noise, and degradation that needed to be improved further by applying an augmentation technique. Fine-tuned TrOCR achieved a character error rate (CER) of 4.98, highlighting good accuracy in detecting scanned receipts [8]. Other work explored variations of the TrOCR model for languages. In [3] explored Decoder-TrOCR for feature extraction in English and Chinese characters. Hybrid integration between TrOCR and ChatBERT [9] achieved better performance, with a CER of 6.03 and a WER of 12.94, but this model was specifically built for English-only. Even though Fine-tuning TrOCR for Spanish [10] revealed limitations, despite achieving great accuracy, it also exhibited forgetting of English recognition ability after fine-tuning and poor performance on multi-line text due to training only on single-line samples.

Further studies have revealed that TrOCR was experimented on varying image effects, only to discover that the model worked highly accurately with black and white text but struggled with color variations and blurred images [11]. Also, edge deployment of Transformer models suffers from accuracy drops, as converting to lightweight formats like TFLite caused a 6.81% performance decrease [12]. However, TrOCR has also been applied for reading information from dashcam footage, such as timestamps and coordinate point, achieving CER of 16 and character recognition rates (CRR) of 84, showing potential in diverse application contexts [13].

In the context of ingredient detection, prior research combining YOLOv5 and OCR [14] has demonstrated promising 98% accuracy, though challenges remain with multi-line OCR and language variations. The other work combining YOLOv5 and EasyOCR [15], reported low accuracy of 12.7% when dealing with complex ingredient layouts. These works are used as references in this study because they collectively highlight both the strengths of TrOCR and YOLO-based pipelines and their current limitations regarding multilingual labels, layout complexity, and deployment robustness in realistic food-packaging conditions.

This study aims to evaluate the performance and robustness of TrOCR and YOLOv8 in extracting ingredient text from food product labels. Particularly, it implements in such conditions as language (Bahasa Indonesia and English), two image backgrounds (Monochrome and RGB), and two text formats (single-line and multi-line) that affect the accuracy of recognition, since there is no recent work combining these variables.

The scope of this study focuses on printed text in food ingredient labels using the TrOCR-based printed model as the core recognizer. The scientific contribution of this work relies on systematically analysing how these visual and linguistic factors jointly affect OCR accuracy, and in demonstrating a unified TrOCR-YOLOv8 framework for multilingual ingredient extraction, which has not been addressed in previous studies that typically focus on a single language, a single layout format. The results of this study are expected to provide insights for improving OCR robustness in multilingual and diverse packaging environments.

2. RESEARCH METHOD

This research was conducted in the form of utilizing YOLOv8 [16], [17], [18] for object detection and TrOCR for text extraction. YOLO, or You Only Look Once, is trained to locate the area of ingredient labels in food packages by annotating each label with a bounding box. Once the bounding box area is detected, the cropped box that contains ingredient text is forwarded to TrOCR for further text extraction.

As TrOCR has a transformer architecture involving both an encoder and a decoder, the encoder is responsible for breaking down and splitting the character into patches, which are then fed into the Visual Transformer (ViT) or CNN backbone. This decoder then processes to generate text tokens from the encoder, to produce readable text outputs from the visual input [19], [20].

This research was carried out by the following steps:

1. Dataset Preparation

   The dataset was compiled from internet sources and camera captures under various lighting conditions. The dataset consisted of 131 images, featuring food packages in both Bahasa Indonesia and English. To assist YOLO for recognizing label ingredient areas, the platform Roboflow helped to draw and annotate the bounding boxes. It was labelled to classify into either "komposisi_ind" or "komposisi_en". Figure 1 and Figure 2 below were samples of the dataset captured by a smartphone camera.

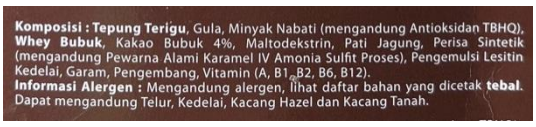


Figure 1. Ingredients in Bahasa Indonesia

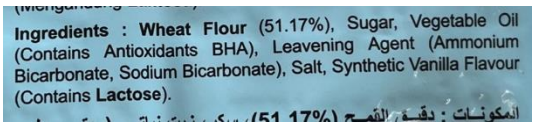


Figure 2. Ingredients in English

2. Data Preprocessing
   Data preprocessing aims to enhance image quality, including resolution, resizing images to 640x640 pixels, reducing image noise, converting to a color format (RGB or Grayscale), adjusting contrast, and normalizing pixel values. These steps are crucial for enhancing both YOLOv8 detection and OCR performance in extraction.

3. Data Augmentation
   Data augmentation is a technique used to increase the size of a dataset and modify it to create various versions of existing data. Moreover, the process aims to enhance the robustness and accuracy [21] of the dataset to be flexible and adaptive in detecting and simulating real-world inputs. Augmentation techniques include rotation, scaling, flipping, adding grain, cropping, adjusting hue and saturation, modifying brightness, blurring, and adding noise. One of the benefits of applying this technique is reducing overfitting by preventing the model from memorizing specific samples, thereby enabling it to handle more diverse inputs.

4. Training Model
   Before training the model, the dataset was split into training, validation, and testing sets with an 85:10:5 ratio. For the training dataset, it was conducted with 100 epochs and 16 batch sizes. YOLOv8 was automatically integrated with TrOCR to detect and extract text from the identified areas.

   The following diagram illustrates the workflow of the pipeline, as shown in Figure 3.

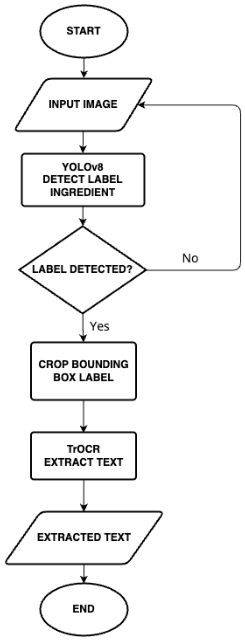


Figure 3. Flowchart

Figure 3 illustrates the overall pipeline of the system for ingredient label recognition and extraction. The process begins by inputting an image, which is then processed using YOLOv8 to detect areas within the image that correspond to ingredient labels. This is captured using a camera phone. If a label is detected, the corresponding bounding box is cropped, focusing on the area of interest. The cropped image is then stored locally and proceeds to TrOCR to extract the text from the area of interest in the label for further analysis to break down the text. Then, the extracted text is displayed to show the compatibility of its performance in feature extraction for characters and words. If no label is detected, the system loops back to allow a new image input to be captured.

In terms of calculating the performance accuracy, this study uses Character Error Rates (CER) and Word Error Rates (WER) as evaluation metrics. These metrics measure the discrepancy between the extracted text and the reference text. The formulas used to calculate these errors are as follows:

$$\text{CER} = \frac{\text{substitutions+deletions+insertions}}{\text{n Characters}} \quad (1)$$

$$\text{WER} = \frac{\text{substitutions+deletions+insertions}}{\text{n Words}} \quad (2)$$

In formulas (1) and (2), substitution means that a character or word in the extracted text is incorrect, similar to the reference. Insertion refers to an extra character or word that has been added to the extracted output. Meanwhile, deletion inherently means the character or word from the reference one is missing in the extraction result. In addition, variable 'n' herein means the total number of characters in reference to the text. A lower error rate is defined as better accuracy [12].

3. RESULTS AND ANALYSIS

3.1. Results and Implementation

TrOCR is inherently be a transformative technology in enhancing productivity without manual reading to text. In this study, a total of 16 testing trials were conducted to evaluate the robustness of TrOCR under various conditions. Three main variables were applied as parameters to measure the accuracy and test the reliability of TrOCR, including two languages (Bahasa Indonesia and English), two types of backgrounds (Monochrome and RGB), and two types of text formatting lines (single-line and multi-line text). These parameters were chosen to simulate real-world scenarios and determine the compatibility and robustness of the OCR model.

The results obtained are presented in the following tables. Table 1 presents two sample inputs and the corresponding extracted outputs by TrOCR, along with the associated error scores. Table 2 outlines the overall error percentages, measured by CER and WER, under each tested condition.

Table 1. Sample Inputs

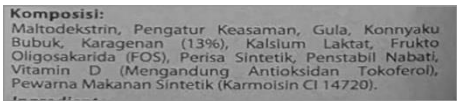
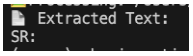
Image Input	Extracted Text	CER / WER (%)
		28.57 / 33.33
		99.21 / 100

Table 2. Overall TrOCR Error Rates

Trials	Language	Line-Type	Background	CER / WER (%)
2	Bahasa Indonesia	Single-Line	Monochrome	29.79 / 51.59
2			RGB	24.77 / 32.15
2		Multi-Line	Monochrome	99.6 / 100
2			RGB	93.93 / 100
2	English	Single-Line	Monochrome	11.99 / 17.15
2			RGB	22.13 / 23.81
2		Multi-Line	Monochrome	89.64 / 100
2			RGB	82.49 / 100

From the results in Table 2, Bahasa Indonesia with a single-line and RGB background showed improvements in both CER and WER compared to the Monochrome Background. The significant improvement was in the word error rate, which was 32.15% for RGB, compared to 51.59% for Monochrome. However, WER was higher than CER, meaning that whole-word recognition was more difficult than individual character recognition. The improvement possibility could occur due to the help of a colored-contrast background in distinguishing characters in Bahasa Indonesia. In contrast, the multi-line format was severely obstructed for Bahasa Indonesia. Despite the fact that, RGB showed a slight improvement over Monochrome, it was still difficult for reading multi-line format, with word errors being 100% for both background types. In this case, there is an implied complete failure at the word-level interpretation.

Contrary to expectations, English in Single-line and Monochrome achieved excellent performance with the lowest CER and WER under 18%. It indicated that English text in single-line and Monochrome format had the best accuracy in the tests, particularly with a CER of 11.99% and a WER of 17.15%. However, the RGB background resulted in higher errors, implying an opposite result compared to the Bahasa Indonesia experiments. This opposite result with the same variables might be caused by language-dependent background contrast, font styles, diacritical marks, or character spacing.

Similar to Bahasa Indonesia with 100% incorrect words, English in a multi-line format performed poorly in both character and word. Merely RGB helped slightly correct the character with 7% better accuracy, rather than 89.64% incorrect characters tested using the Monochrome format. This highlighted a consistent

limitation, regardless of the language, which TrOCR lacked in line segmentation or struggled to maintain context across lines.

By observing the first sample in Table 1, which contains long words in a single-line format, notice that it has a total of nine words. However, the extracted text was only able to read seven words correctly with two deletions. On average, the TrOCR model successfully extracted six out of eight long words per sample, indicating the other two words were missing or misrecognized during the recognition process. This implies that the TrOCR model struggled to accurately extract more than 6 words from a single line input. These issues might be caused by the tokens reaching their maximum limitations in generating or extracting text. Transformer architecture with self-attention breaks the sentence into patches to comprehend the entire sentence. It has limitations in understanding long and complex passages, which may cause confusion and a loss of focus over time when recognizing long words or lines. This reduced its ability to focus on words toward the end of a line.

TrOCR's underperformance in a multi-line format was likely due to its limited capability in understanding spatial layout. Multi-line recognition required the model to recognize text that contains vertical and horizontal spacing simultaneously. This ability was vital for the OCR model to determine which line to extract at a time and when to move to the next line. By ignoring this line segmentation, the model either extracted text incorrectly or deleted some parts completely.

3.2. Discussion

The TrOCR-YOLOv8 is proficient in detecting and cropping regions of ingredient labels, as indicated by the experimental results. However, its text recognition performance is highly sensitive to layout complexity, text length, and visual presentation: single-line English text on Monochrome backgrounds achieved the lowest error rates, while performance sharply declined for longer lines and was especially poor for multi-line text. CER and WER reached up to 99.6% and 100% for Bahasa Indonesia in Monochrome backgrounds, indicating limitations in attention span and spatial layout comprehension. These results extend previous work by methodically demonstrating how language (Bahasa Indonesia vs. English), background (Monochrome vs. RGB), and text formats (single line vs. multi-line) jointly constrain TrOCR in realistic food-label scenarios. Based on these findings, the proposed method is best suited for its implementations involving short, single-line text, such as license plates, product labels, timestamps, signs, or scene text, particularly where text layout is simple. While highlighting the need for future OCR systems to incorporate explicit line segmentation and layout-aware modeling to handle dense, multilingual ingredient lists more reliably.

4. CONCLUSION

In conclusion, YOLOv8 demonstrated a good capability in detecting and cropping the bounding box based on its label ingredient areas. However, the TrOCR base printed model faced challenges in feature extraction. One major challenge was poor performance on multi-line text formatting, as evidenced by extremely high error rates that reached up to 99.6% CER and 100% WER in Bahasa Indonesia with a Monochrome background. Despite a single line that could be more readable by a transformer, the model still struggled when the text exceeded six words on average. In this case, the model often failed to extract the entire text, resulting in missing characters and words. The other effects, like image quality and background color, played a significant role in the accuracy of feature extraction.

REFERENCES

[1] M. Li et al., "TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models," 2023. [Online]. Available: www.aaai.org

[2] C. Gunasekara, Z. Hamel, F. Du, and C. Baillie, "TokenOCR: An Attention Based Foundational Model for Intelligent Optical Character Recognition," in International Conference on Pattern Recognition Applications and Methods, Science and Technology Publications, Lda, 2025, pp. 151–158. doi: 10.5220/0013340100003905.

[3] M. Fujitake, "DTrOCR: Decoder-only Transformer for Optical Character Recognition."

[4] L. Beerens and D. J. Higham, "Vulnerability Analysis of Transformer-based Optical Character Recognition to Adversarial Attacks," Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.17128

[5] P. B. Ströbel, S. Clematide, M. Volk, and T. Hodel, "Transformer-based HTR for Historical Documents," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.11008

[6] W. Khallouli, M. S. Uddin, A. Sousa-Poza, J. Li, and S. Kovacic, "Leveraging Transformer-Based OCR Model with Generative Data Augmentation for Engineering Document Recognition †," Electronics (Switzerland), vol. 14, no. 1, Jan. 2025, doi: 10.3390/electronics14010005.

[7] A. Mortadi et al., "ALNASIKH: An Arabic OCR System Based on Transformers," in 3rd International Mobile, Intelligent, and Ubiquitous Computing Conference, MIUCC 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 74–81. doi: 10.1109/MIUCC58832.2023.10278370.

[8] H. Zhang, E. Whittaker, and I. Kitagishi, "Extending TrOCR for Text Localization-Free OCR of Full-Page Scanned Receipt Images."

[9] Y.-H. Chen and P. B. Ströbel, "TrOCR Meets Language Models: An End-to-End Post-correction Approach," 2024, pp. 12–26. doi: 10.1007/978-3-031-70645-5_2.
[10] F. Lauar and V. Laurent, "Spanish TrOCR: Leveraging Transfer Learning for Language Adaptation," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.06950
[11] R. L. Zhang, "A Comprehensive Evaluation of TrOCR with Varying Image Effects," 2024.
[12] R. Ahmed, N. Shabbir, M. W. Raza, A. Zeb, and H. Elahi, "Evaluation of Model Degradation in PaddleOCR, UltOCR, and TrOCR Across Baseline and TensorFlow Lite Environments," in 6th International Conference on Robotics and Automation in Industry, ICRAI 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICRAI62391.2024.10894257.
[13] S. Chandrasekaran, K. I. Ramachandran, S. Adarsh, and B. B. Nair, "Graphical Abstract Transformers for Dashcam: Extraction of Timestamps and GPS." [Online]. Available: https://ssrn.com/abstract=4975821
[14] S. Tarannum, M. S. Jalal, and M. N. Huda, "HALALCheck: A Multi-Faceted Approach for Intelligent Halal Packaged Food Recognition and Analysis," IEEE Access, vol. 12, pp. 28462–28474, 2024, doi: 10.1109/ACCESS.2024.3367983.
[15] R. Farokhynia and M. Krikeb, "Simultaneous Detection and Validation of Multiple Ingredients on Product Packages: An Automated Approach," 2024.
[16] M. Hussain, "YOLOv5, YOLOv8 and YOLOv10: The Go-To Detectors for Real-time Vision," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.02988
[17] M. Sohan, T. Sai Ram, and Ch. V. Rami Reddy, "A Review on YOLOv8 and Its Advancements," 2024, pp. 529–545. doi: 10.1007/978-981-99-7962-2_39.
[18] R. Varghese and M. Sambath, "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," in 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems, ADICS 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ADICS58448.2024.10533619.
[19] D. Chang and Y. Li, "DLoRA-TrOCR: Mixed Text Mode Optical Character Recognition Based On Transformer," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.12734
[20] C. P. Vossos, "Handwritten Optical Character Recognition," 2024.
[21] T. B. Pun, A. Neupane, R. Koech, and K. Walsh, "Detection and counting of root-knot nematodes using YOLO models with mosaic augmentation," Biosens Bioelectron X, vol. 15, no. September, p. 100407, 2023, doi: 10.1016/j.biosx.2023.100407.

BIBLIOGRAPHY OF AUTHORS

Charina Mutiara Chairunnisa, was born in Sungai Penuh. She earned a Bachelor of Applied Electronics Engineering degree from the Department of Electrical Engineering at Politeknik Negeri Sriwijaya. Her research interests focus on computer vision and deep learning.

Nyayu Latifah Husni, is a lecturer of Electrical Engineering, Department of Electrical Engineering, at Politeknik Negeri Sriwijaya.

RD. Kusumanto, is a lecturer of Electrical Engineering, Department of Electrical Engineering, at Politeknik Negeri Sriwijaya.