International Journal of Engineering and Computer Science Applications (IJECSA)
Vol. 5, No. 2, September 2026, pp. 55∼60
ISSN: 2828-5611
DOI: 10.30812/IJECSA.v5i2.6620

Enhanced Convolutional Recurrent Neural Network for Vertical Japanese Text Recognition

Ronny, Sandy Kosasi, David
STMIK Pontianak, Pontianak, Indonesia

Article Info

Article history:

Received July 1, 2025
Revised July 24, 2025
Accepted September 30, 2025

Keywords:

Attention Mechanism
CRNN
OCR
Residual Connection
Vertical Text

ABSTRACT

Optical Character Recognition (OCR) plays a crucial role in text digitization, yet most existing research predominantly focuses on horizontal English or Latin text, leaving vertical Japanese text largely underexplored. This research addresses that gap by proposing an enhanced Convolutional Recurrent Neural Network (CRNN) model tailored for Japanese vertical text recognition. The novelty of this work lies in integrating residual connections and attention mechanisms into the CRNN, along with a custom dataset that includes both Japanese characters and 30% Latin characters, simulating real-world scenarios such as manga. Experimental results show that the proposed model achieves a Character Error Rate (CER) of 0.363% in only 10 epochs, compared to the baseline CRNN, which reaches 2.084% CER after 100 epochs, demonstrating both faster convergence and improved accuracy. Furthermore, evaluation on manga screenshots from the Manga109 dataset highlights the model's practical potential, while also revealing new challenges related to irregular vertical spacing and font variability. These findings indicate that the proposed method significantly advances vertical Japanese text recognition and offers a foundation for future improvements in text recognition for manga and similar media.

Corresponding Author:

Ronny,
STMIK Pontianak, Pontianak, Indonesia
Email: zreactives@gmail.com

How to Cite: R. Ronny, S. Kosasi, and D. David, "Enhanced Convolutional Recurrent Neural Network for Vertical Japanese Text Recognition", *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 5, no. 2, pp. 55–60, Sept. 2026. doi: 10.30812/ijecsa.v5i2.6620.

1. INTRODUCTION

Japanese text, written from top to bottom and read from right to left, is widely used in printed media such as manga, newspapers, and books. This orientation differs significantly from the horizontal writing systems that dominate most global languages, such as the Latin script. As a result, the majority of Optical Character Recognition (OCR) systems are designed and optimized for horizontal text, which poses challenges for recognizing vertically oriented scripts. These are more challenging when texts are obtained from sources such as scanned documents or photos, which often introduce distortions, blurring, font variations, and noise [1, 2]. To achieve high recognition accuracy, OCR systems must not only be robust to image degradation but also capable of modeling the sequential structure of vertical character layouts [3]. The importance of addressing these challenges is highlighted in studies of furigana detection, where removing furigana (small annotations above kanji characters) improved OCR performance by up to 5% on the Manga109 dataset [1]. This demonstrates that additional textual artifacts can interfere with recognition accuracy. Moreover, benchmark studies on cross-lingual OCR show that recognition performance on Asian languages, including Japanese, Korean, and Arabic, is significantly lower compared to Latin-based scripts, with Japanese text achieving the lowest performance [1, 2]. Orientation remains a key factor in this performance gap, with multi-oriented text recognition resulting in accuracy drops of up to 23% [2]. Similarly, Yu et al. reported that a baseline Convolutional Recurrent Neural Network (CRNN) achieved only 51.97% accuracy on vertical Chinese text recognition (VCTR), underscoring the need for more robust architectures [3]. Beyond traditional CRNN architectures, alternative designs, such as Convolutional Character Networks, have been proposed to improve recognition by directly modeling character-level representations using convolutional structures [4]. However, such models are still primarily evaluated on horizontally oriented scripts, and their effectiveness on vertical Japanese text remains limited. This highlights the need for specialized OCR architectures that can robustly handle vertical orientation while maintaining efficiency. Practical, lightweight OCR systems like PP-OCRv3 also report notable gains by upgrading both detection and recognition (e.g., the SVTR recognizer, residual-attention modules, more potent distillation and augmentation), suggesting that attention- and residual-based designs help in real-world multi-orientation scenarios [4]. Complementary architectural choices, such as multi-scale fusion CRNNs, expand the receptive field across character sizes and layouts, improving recognition in complex backgrounds [5]. Regularization choices (e.g., label smoothing) and lightweight language modeling further enhance CRNN generalization on irregular and noisy text conditions typical of scanned/photographed vertical materials [6].

The CRNN framework, combining convolutional layers for feature extraction and recurrent layers for sequence modeling, has been widely adopted as a baseline for OCR tasks [7]. Despite its effectiveness, CRNN alone struggles to capture complex contextual dependencies in challenging settings such as vertical Japanese text. Recent works have shown that attention mechanisms can enhance OCR models by dynamically focusing on informative regions of the input sequence, thereby significantly improving recognition in complex scripts [8, 9]. Furthermore, the integration of residual connections has been shown to stabilize training and improve feature propagation in deep networks, making them more resilient to noise and degradation [10]. Several multilingual OCR studies have highlighted the accuracy gap between Latin and Asian scripts, underscoring the limitations of current general-purpose models and the need for domain-specific architectures [11, 12]. Benchmark datasets that include vertical or multi-oriented scripts, such as those for Chinese and Japanese, have been instrumental in evaluating the robustness of OCR models in real-world scenarios [13]. Together, these findings indicate that vertical Japanese text recognition remains an open and challenging problem that demands specialized model designs.

In this research, we introduce a modified CRNN architecture tailored for vertical Japanese text recognition. The proposed model incorporates four convolutional layers in the CNN backbone, followed by two stacked bidirectional LSTM layers with dual hidden states. To enhance contextual modeling, we integrate a multi-head attention mechanism, enabling the network to capture long-range dependencies across vertical sequences. Additionally, residual connections are employed to ensure stable gradient flow and improve feature learning in deeper layers. The main contributions of this research are modifying CRNN architecture with multi-head attention and residual connections to effectively capture vertical contextual dependencies, integration into a real-time OCR pipeline capable of recognizing vertical Japanese text directly from offline sources such as scanned documents or photographs, and comprehensive evaluation using Character Error Rate (CER) on a manually adapted vertical text dataset derived from the ETL Character Database, with comparisons against a baseline CRNN implementation. By addressing the limitations of existing OCR models, this study advances robust recognition of vertically oriented Japanese text and provides insights applicable to other complex writing systems.

2. RESEARCH METHOD

This study employs an experimental research methodology to develop and evaluate a Convolutional Recurrent Neural Network (CRNN) for vertical Japanese text recognition. The research consists of steps including dataset collecting and preprocessing, model

architecture design, training procedure, and evaluation, which is illustrated in Figure 1.

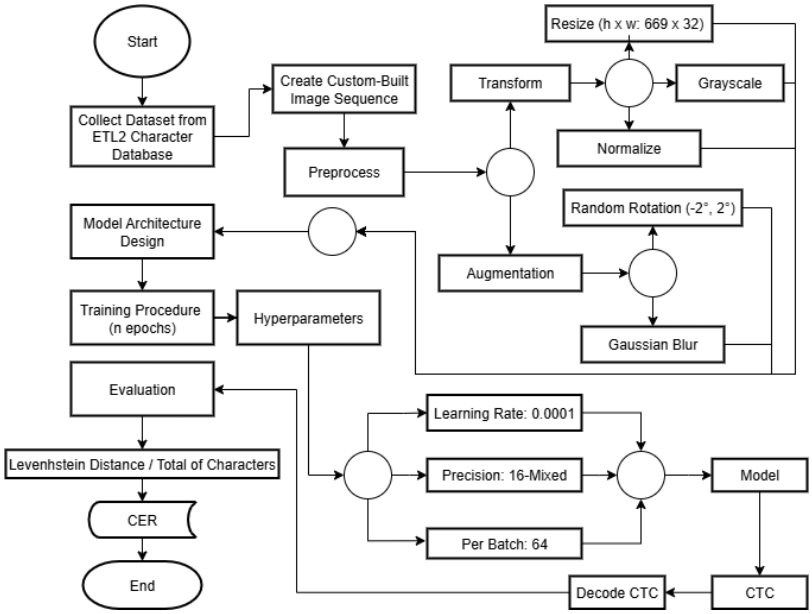

Figure 1. Research Flow Flowchart

The primary dataset used in this research consists of scanned Japanese characters from the ETL2 Character Database, which contains grayscale Images of characters from Japanese newspapers. The dataset includes 2184 characters, consisting of hiragana, katakana, alphanumeric characters, symbols, and kanji. To form training samples, characters were merged into vertical sequences representing sentences. Preprocessing was performed using the torchvision transforms module by resizing images to h×w (669×32) for consistency, converting to grayscale (1 channel), and normalizing with a mean and std of 0.5 to improve performance and accuracy [14]. The image is then augmented to simulate noise in a real-life scenario, like random rotation, which is applied with degrees of -2 to 2, and size expanded to make sure there are no erased parts, which tends to be very destructive towards the model, and then a Gaussian blur to simulate a low-quality scan to help increase robustness in the model. Each augmentation was designed to have a 20% chance of occurring, allowing the model to learn from the original image while maintaining robustness.

The model's baseline is CRNN, comprising a CNN backbone for feature extraction, stacked BiLSTMs for sequence modeling, and a CTC transcription layer [15]. To better capture vertical text dependencies, we introduce two modifications: a multi-head attention mechanism to focus on multiple character relationships, and residual connections to stabilize training and mitigate vanishing gradients in deeper layers [16, 17]. The random rotation augmentation with the expand feature also makes the image width dynamic (which is supposed to be fixed) and could cause the model to fail. To address this issue, augmentation was applied first, and then the image was resized to ensure a fixed width. The overall architecture is illustrated in Table 1.

Table 1. Model Architecture Configuration. "k", "s", and "p" stand for kernel size, stride, and padding, respectively.

Type	Configurations
Input	Grayscale, Dynamic W x H image
Convolution Set	Maps: 128, k: 3 x 3, s: 1, p: 1, Including BatchNorm and ReLU
Convolution Set with 2x2 pool	Maps: 128, k: 3 x 3, s: 1, p: 1, Including BatchNorm and ReLU
Convolution Set	Maps: 256, k: 3 x 3, s: 1, p: 1, Including BatchNorm and ReLU
Convolution Set with 2x2 pool	Maps: 256, k: 3 x 3, s: 1, p: 1, Including BatchNorm and ReLU
Fully Connected Layer	Output: 256
LSTM 1 with layer normalization	Hidden Units: 256, Num Layer: 2, Bidirectional
Multi-Head Attention	Heads = 8
LSTM 2 with layer normalization	Hidden Units: 256, Num Layer: 2, Bidirectional
Residual Connection	LSTM 1 + Multi-Head Attention + LSTM 2
Fully Connected Layer	Output: Number of Class (2168 + 1 (blank class))

The model is trained using the CTC (Connectionist Temporal Classification) loss function, suitable for sequence labeling [18]. The model is optimized using the Adam Optimizer (learning rate = 0.0001). Mixed-precision training and a batch size of 64 are used to improve computational efficiency. These experiments are carried out on a laptop with a 2.3 GHz Intel(R) Core(TM) i7-12650H (16 CPU cores), 40 GB of RAM, and an NVIDIA RTX 3060 GPU (6GB VRAM). Model Accuracy is measured using the Character Error Rate (CER), which is the ratio of insertions, deletions, and substitutions to the ground truth length [9]. To measure the CER in this experiment, the Levenshtein distance, which quantifies word similarity, was integrated with OCR to obtain the CER [19].

3. RESULT AND ANALYSIS

3.1. Experimental Setup

The proposed CRNN model was evaluated on a custom-built dataset consisting of Japanese vertical text extracted from the ETL Character Database. To ensure consistency, all images were fixed to a width of 32 and a height of 669. The dataset was constructed in several partitions: 8000 training samples, an additional 4000 training samples containing 30% Latin (alphanumeric) characters to increase script diversity, 1500 validation samples, and a 1500 validation dataset with a 30% Latin ratio. For optimization, the Adam optimizer was employed with an initial learning rate of 1e-4. The loss function chosen for sequence recognition was the Connectionist Temporal Classification loss, as it is well-suited for handling unsegmented sequence data, which will be helpful in this experiment.

3.2. Training Result

To evaluate the effectiveness of the proposed architecture, experiments were carried out without data augmentation. Preliminary findings indicated that when trained on those data, our CRNN model consistently achieved lower Character Error Rates (CER) and converged more rapidly compared to the baseline CRNN. The performance comparison between the baseline CRNN model and our proposed approach is illustrated in Figure 2. The learning rate was kept constant across both configurations to ensure a fair comparison.

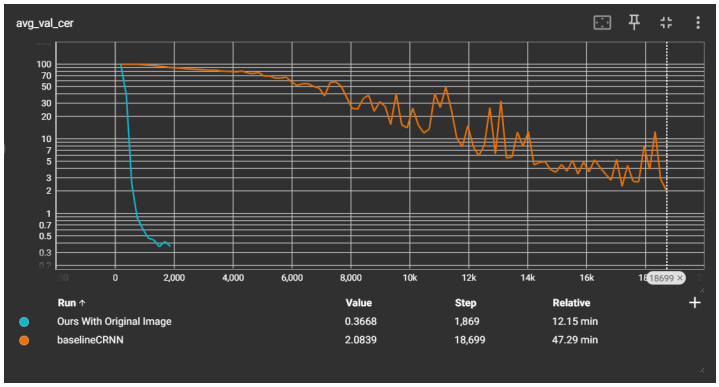

Figure 2. Comparison Graph of Our CRNN Model and Baseline CRNN using Tensorboard

From the graph, it can be observed that the validation CER of our proposed CRNN decreases much more sharply in the initial epochs and stabilizes earlier, reflecting a faster learning process. Specifically, the lowest CER achieved by our model was 0.363%, obtained at the 7th epoch. By contrast, the baseline CRNN attained a lowest CER of 2.084%, which only appeared at the final (100th) epoch. This corresponds to an 82.58% relative improvement in CER performance by our proposed model compared to the baseline. These findings demonstrate two essential aspects: the modified architecture not only improves recognition accuracy but also significantly accelerates convergence, and the efficiency of our approach enables superior results in fewer training epochs, thereby reducing the overall computational cost of our training procedure.

3.3. Evaluation on Manga109 Dataset

To further assess the robustness of the proposed CRNN model in real-world scenarios, we conducted an additional evaluation on vertical-text samples extracted from the Manga109 dataset [10, 20]. We created a custom test set by manually capturing screenshots of vertical Japanese text regions from selected manga volumes. All captured text regions were preprocessed to match the same input format described in Section 3.1, with the height fixed to 32 pixels while preserving the aspect ratio for the width, rather than a fixed

width of 669 pixels. Importantly, no fine-tuning was performed on this dataset; the models trained solely on the ETL-derived dataset were applied directly. This allows us to evaluate the cross-domain generalization of the models when faced with more complex and noisy backgrounds. The evaluation results are shown in Figure 3.

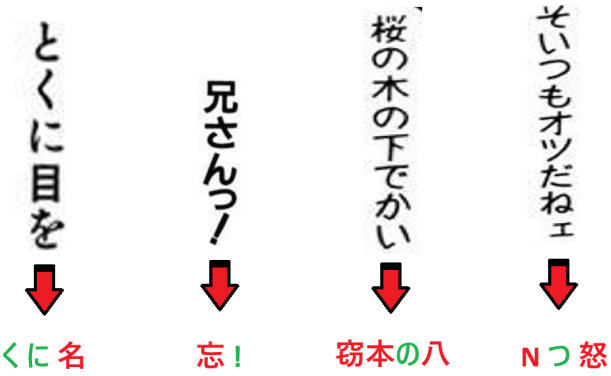

Figure 3. Evaluation Result of Manga109 Dialog Screenshot. Green and Red colors represent Correct and Wrong, respectively.

4. CONCLUSION

In this study, we proposed a modified Convolutional Recurrent Neural Network (CRNN) architecture for vertical Japanese text recognition and evaluated its performance against a baseline CRNN model. Experimental results on ETL-derived datasets demonstrated that our model consistently achieved faster convergence and a significantly lower Character Error Rate (CER) than the baseline. Notably, the proposed CRNN reached a minimum CER of 0.363% in only seven epochs, whereas the baseline required 100 epochs to achieve a minimum CER of 2.084%. This corresponds to an 82.58% improvement in error reduction, indicating the effectiveness of the proposed modifications. Further evaluation on manga screenshots from the Manga109 dataset revealed the challenges of applying the model to real-world text recognition tasks. The primary sources of recognition errors were variations in vertical spacing between characters and the use of diverse fonts, as well as the background noise inherent in manga. Although the absolute accuracy decreased in this domain, the proposed CRNN still outperformed the baseline, suggesting that the architectural improvements contribute to better generalization under more complex conditions. Overall, the findings confirm that the proposed CRNN architecture provides a more efficient and accurate solution for vertical Japanese text recognition. However, to bridge the gap between controlled datasets and real-world applications, future research should explore domain adaptation strategies, such as fine-tuning on manga-specific text, augmenting training data with varied spacing and fonts. These directions are expected to further enhance the robustness of OCR systems in practical applications such as instant manga translation.

REFERENCES

[1] N. K. Bjerregaard, V. Cheplygina, and S. Heinrich. “Detection of Furigana Text in Images.” version 1, pre-published.

[2] Z. Yang et al. “CC-OCR: A Comprehensive and Challenging OCR Benchmark for Evaluating Large Multimodal Models in Literacy.” arXiv: 2412.02210 [cs.CV], pre-published.

[3] H. Yu et al. “Orientation-Independent Chinese Text Recognition in Scene Images.” arXiv: 2309.01081 [cs.CV], pre-published.

[4] C. Li et al. “PP-OCRv3: More Attempts for the Improvement of Ultra Lightweight OCR System.” arXiv: 2206.03001 [cs.CV], pre-published.

[5] L. Zou et al., “Text Recognition Model Based on Multi-Scale Fusion CRNN,” *Sensors*, vol. 23, no. 16, p. 7034, Aug. 8, 2023. DOI: 10.3390/s23167034

[6] W. Yu, M. Ibrayim, and A. Hamdulla, “Scene Text Recognition Based on Improved CRNN,” *Information*, vol. 14, no. 7, p. 369, Jun. 28, 2023. DOI: 10.3390/info14070369

[7] X. Chen et al., “Text Recognition in the Wild: A Survey,” *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–35, Mar. 31, 2022. DOI: 10.1145/3440756

[8] T. Clanuwat, A. Lamb, and A. Kitamoto, "KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia: IEEE, Sep. 2019, pp. 607–614. DOI: 10.1109/ICDAR.2019.00103

[9] R. Hinami et al., "Towards Fully Automated Manga Translation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, pp. 12 998–13 008, May 18, 2021. DOI: 10.1609/aaai.v35i14.17537

[10] K. Aizawa et al., "Building a Manga Dataset "Manga109" With Annotations for Multimedia Applications," *IEEE MultiMedia*, vol. 27, no. 2, pp. 8–18, Apr. 1, 2020. DOI: 10.1109/MMUL.2020.2987895

[11] P. Dai and X. Cao. "Comprehensive Studies for Arbitrary-shape Scene Text Detection." arXiv: 2107.11800 [cs.CV], pre-published.

[12] J. Baek et al., "What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 4714–4722. DOI: 10.1109/ICCV.2019.00481

[13] J. Ye et al., "TextFuseNet: Scene Text Detection with Richer Fused Features," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, Jul. 2020, pp. 516–522. DOI: 10.24963/ijcai.2020/72

[14] X. Pei et al., "Robustness of machine learning to color, size change, normalization, and image enhancement on micrograph datasets with large sample differences," *Materials & Design*, vol. 232, p. 112 086, Aug. 2023. DOI: 10.1016/j.matdes.2023.112086

[15] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, Nov. 1, 2017. DOI: 10.1109/TPAMI.2016.2646371

[16] A. Vaswani et al. "Attention Is All You Need." version 7, pre-published.

[17] K. He et al., "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90

[18] W. Hu et al., "GTC: Guided Training of CTC towards Efficient and Accurate Scene Text Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11 005–11 012, Apr. 3, 2020. DOI: 10.1609/aaai.v34i07.6735

[19] M. G. Pradana et al., "Levenshtein Distance Algorithm in Javanese Character Translation Machine Based on Optical Character Recognition," *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 4, p. 1411, Jul. 30, 2025. DOI: 10.62527/joiv.9.4.3151

[20] Y. Matsui et al., "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21 811–21 838, Oct. 1, 2017. DOI: 10.1007/s11042-016-4020-z