International Journal of Future Engineering Innovations

A Model Selection Framework Based on Writing Quality Levels for Image-Based Text Recognition

Phuong Nguyen Thi Thanh [1], Khuong Pham Phu [2], Trinh Tran Van [3], Hieu Ngo Van [4*]
[1-2, 4] School of Computer Science and Artificial Intelligence, Duy Tan University, Danang, Vietnam
[3] International School, Duy Tan University, Danang, Vietnam

* Corresponding Author: **Hieu Ngo Van**

Article Info

ISSN (online): 3049-1215
Impact Factor (RSIF): 8.25
Volume: 03
Issue: 03
May-June 2026
Received: 26-02-2026
Accepted: 27-03-2026
Published: 30-04-2026
Page No: 20-25

Abstract
This study proposes a writing-quality-aware model selection framework for image-based text recognition to address the heterogeneity of real-world data. Three representative models, including CNN, ResNet combined with BiLSTM-CTC, and TrOCR, are evaluated across three data levels: printed text, clean handwriting, and poor handwriting. Experimental results show that model performance decreases as data quality deteriorates, while TrOCR consistently achieves the best robustness and accuracy. Based on these observations, the proposed framework consists of two stages: estimating the quality of input images and selecting the most suitable model for each quality level. Validation on independent datasets demonstrates that the framework maintains stable performance and improves overall accuracy compared to using a single fixed model, highlighting the effectiveness of quality-aware model selection for real-world OCR applications.

DOI: https://doi.org/10.54660/IJFEI.2026.3.3.20-25

Keywords: Image-Based Text Recognition, Writing Quality, Model Selection, OCR, Deep Learning

1. Introduction

Text recognition in document images [1] is an important problem in the field of computer vision [2], serving as a fundamental component for many real-world applications such as document digitization, automatic information extraction, and content-based search systems. In recent years, the development of deep learning models has significantly improved the performance of text recognition systems.

However, the effectiveness of these models heavily depends on the quality of input data. In practice, document images are often heterogeneous, including clean printed text, readable handwriting, and poor handwriting affected by noise, blur, or geometric distortions. This makes evaluating models on a single type of data insufficient to fully reflect their real-world performance.

In addition, most existing studies focus on improving model architectures to enhance overall accuracy, while paying limited attention to performance variations across different data quality levels. This limitation reduces the applicability of models in real-world systems that must handle diverse types of documents.

Motivated by this issue, this study proposes a framework for evaluating and selecting text recognition models based on writing quality levels in document images, aiming to analyze and compare model performance under different data conditions. The proposed framework not only enables more comprehensive evaluation but also supports selecting appropriate models for specific application scenarios.

The main contributions of this study are as follows: (i) proposing a data categorization framework with three writing quality levels, including printed text, clean handwriting, and poor handwriting; (ii) building an experimental system based on three representative models, including a basic CNN [3], ResNet combined with BiLSTM-CTC [4], and the Transformer-based TrOCR [5]; and (iii) introducing a cross-dataset validation mechanism to evaluate the generalization ability and robustness of the proposed framework under different independent data conditions.

2. Related Work

In recent years, the task of text recognition in images has undergone a clear shift from traditional CNN (baseline) approaches to Transformer-based architectures. This transition is driven by the ability of Transformers to model global contextual relationships while reducing the reliance on handcrafted components in the OCR pipeline.

Li *et al*. (2021) [6] proposed TrOCR, which combines a Vision Transformer (ViT) with a Transformer decoder to directly convert images into character sequences. This approach eliminates traditional handcrafted feature extraction steps and achieves strong performance on multiple standard OCR benchmarks.

Subsequently, Bautista *et al*. (2022) [7] introduced PARSeq, an autoregressive Transformer architecture for scene text recognition. The model leverages attention mechanisms to better capture character-level dependencies within sequences, resulting in improved accuracy across several benchmark datasets.

More recently, Blecher *et al*. (2023) [8] proposed Nougat, a Transformer-based model designed for converting complex documents such as PDFs and scientific document images into structured text. The model demonstrates strong performance in handling documents with complex layouts, expanding OCR applications into academic and technical domains.

Although recent studies have achieved impressive results, most of them primarily focus on improving model architectures to enhance overall accuracy. However, these models are typically evaluated on relatively homogeneous datasets or under fixed data conditions, without fully considering variations in writing quality such as printed text, clean handwriting, and poor handwriting.

This leads to an important research gap: there is no unified evaluation framework based on writing quality levels to systematically analyze model performance under different data conditions. Therefore, developing a quality-aware evaluation framework is necessary to more accurately reflect the real-world applicability of text recognition models.

3. Experimental Comparison of Text Recognition Models Across Different Writing Quality Levels

In this section, we evaluate and compare the performance of text recognition models on three writing quality levels, including printed text, clean handwriting, and poor handwriting. The experiments focus on analyzing the impact of input data quality on the recognition ability of each model, thereby highlighting their stability and generalization capability under different real-world data conditions.

3.1. Training Models and Datasets

In the initial experiments, this study selects three representative models corresponding to different approaches in image-based text recognition. The first model is a traditional CNN, serving as a baseline to evaluate basic performance based on local feature extraction. The second model is a ResNet combined with BiLSTM and CTC loss, representing deep learning approaches that integrate feature extraction with sequence modeling. The third model is TrOCR, representing a modern Transformer-based approach for direct image-to-text recognition.

Regarding datasets, the study uses three datasets corresponding to different writing quality levels to reflect real-world conditions.

The first dataset is IIIT 5K-Words [9], representing printed text with clear character structures and stable layouts. The second dataset is the IAM Handwriting Database [10], which contains high-quality handwritten text with relatively clear and readable characters. The third dataset is a handwritten dataset collected from students under real-world conditions, which has lower quality due to noise, blur, and geometric distortions. To better simulate realistic scenarios, this dataset is further augmented using techniques such as blurring, noise injection, and geometric transformations.

All models are trained and evaluated under the same data structure across the three quality levels to ensure a fair comparison and to accurately reflect the data processing capability of each method.

3.2. Experimental Setup and Evaluation Protocol

3.2.1. Experimental setup

In this study, the dataset is organized into three subsets corresponding to different writing quality levels, including printed text, clean handwriting, and poor handwriting. Each subset represents a different difficulty level in the image-based text recognition task.

To enable the learning of generalizable features, the training set is constructed by combining all training data from the three writing quality levels into a single unified dataset. This approach allows the models to learn diverse representations under different data conditions within a single training process.

In contrast, the test set is separated according to writing quality, resulting in three independent test subsets corresponding to printed text, clean handwriting, and poor handwriting. Each test subset contains data from only one quality level, ensuring a fair evaluation of the model's generalization ability under specific conditions.

3.2.2. Evaluation protocol

After training on the combined dataset, all models are evaluated on each test subset separately. For each input image, the model generates a predicted character sequence, which is then compared with the ground truth to compute evaluation metrics.

The metrics used in this study include Accuracy [11], Character Error Rate (CER) [12], and Word Error Rate (WER) [13], which comprehensively reflect recognition performance at both character and word levels.

The results are recorded separately for each model across all test conditions and summarized in comparative tables for further analysis. All models are evaluated under the same training settings and using the same metrics to ensure a fair comparison.

3.3. Experimental Results and Analysis

In this section, we present the experimental results to evaluate the performance of text recognition models across three writing quality levels, including printed text, clean handwriting, and poor handwriting. The results are analyzed using standard evaluation metrics such as Accuracy, CER, and WER, thereby highlighting differences in data processing capability among the models under heterogeneous input conditions. This also provides insights into the stability and generalization ability of the methods when dealing with increasingly difficult data.

3.3.1. Experimental results

The table below presents a performance comparison of three representative models across three different data levels, clearly reflecting the changes in model performance corresponding to different writing quality conditions.

Table 1: Experimental Results Across Writing Quality Levels

Model	Data Level	Accuracy (%)	CER (%)	WER (%)
CNN (baseline)	Printed text	86%	18%	24%
CNN (baseline)	Clean handwriting	74%	29%	38%
CNN (baseline)	Poor handwriting	58%	45%	61%
ResNet + BiLSTM + CTC	Printed text	92%	11%	17%
ResNet + BiLSTM + CTC	Clean handwriting	84%	21%	30%
ResNet + BiLSTM + CTC	Poor handwriting	70%	34%	48%
TrOCR	Printed text	96%	6%	10%
TrOCR	Clean handwriting	90%	14%	21%
TrOCR	Poor handwriting	82%	25%	36%

3.3.2. Analysis of Results

Based on the experimental results in the table 1, it can be observed that the performance of all models decreases as the writing quality degrades from printed text to clean handwriting and and poor handwriting. At the printed text level, all models achieve high performance, with TrOCR obtaining the best results (96% accuracy), followed by ResNet + BiLSTM + CTC (92%) and CNN (86%). This indicates that when the input data is clean and less noisy, all models perform well, although modern architectures still demonstrate clear advantages.

When moving to clean handwriting data, performance differences among models become more evident. CNN drops to 74%, while ResNet + BiLSTM + CTC achieves 84% and TrOCR maintains a high performance of 90%. This shows that models capable of capturing deeper representations and contextual information (such as ResNet-BiLSTM and Transformer-based models) are more effective than traditional CNNs under more complex data conditions.

At the poor handwriting level, the performance gap becomes most significant. CNN decreases sharply to 58%, ResNet + BiLSTM + CTC reaches 70%, while TrOCR still achieves the best performance at 82%. Similarly, its error metrics (CER and WER) are also the lowest (25% and 36%), indicating greater robustness in noisy and heavily distorted conditions.

Overall, the results show that CNN is highly sensitive to data quality degradation, while ResNet + BiLSTM + CTC improves performance by modeling sequential dependencies. However, TrOCR outperforms both due to its ability to learn global representations through attention mechanisms, enabling stable performance across all data conditions. This confirms that Transformer-based models are more adaptable to heterogeneous data in text recognition tasks

3.3.3. Comparison chart

The figure below illustrates the performance trend (Accuracy) of three models across three different writing quality levels.

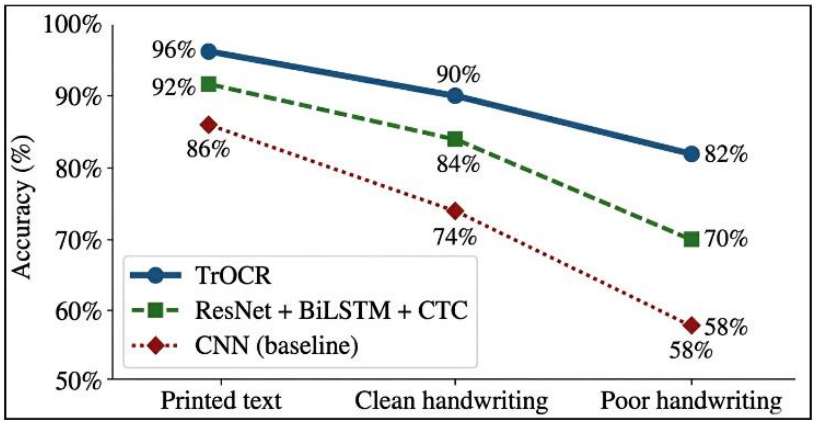

Fig 1: Model performance comparison chart.

Specifically, the x-axis represents the data conditions, including Printed text, Clean handwriting, and Poor handwriting, while the y-axis shows the Accuracy metric. The three curves correspond to CNN, ResNet + BiLSTM + CTC, and TrOCR.

The results indicate that all models experience a performance decline as the difficulty of the input data increases. However, the degree of degradation varies among models. The CNN (baseline) shows the most significant drop, especially when moving from printed text to poor handwriting. ResNet + BiLSTM + CTC demonstrates more stable performance due to its ability to combine deep feature extraction with sequential modeling.

In contrast, TrOCR maintains the highest and most stable performance across all three levels, highlighting the advantage of Transformer-based architectures in capturing global context and handling noisy inputs. This confirms the superior generalization ability of Transformer-based models in text recognition tasks.

4. Proposed Framework For Evaluating Text Quality Levels

Experimental results show that the performance of text recognition models varies depending on the quality level of text in document images. CNN-based models are suitable for printed text but their performance significantly decreases when handling handwritten text. The ResNet combined with BiLSTM and CTC provides better stability across different data conditions. Meanwhile, TrOCR achieves the highest and most stable performance across all quality levels.

These findings indicate that using a single fixed model for all cases is not optimal. Therefore, this study proposes a model selection framework based on text quality levels to improve recognition performance under varying data conditions.

4.1. Operating Principle of the Proposed Framework

The proposed framework is designed as a two-stage decision process to select an appropriate recognition model based on the input image quality.

4.1.1. Input image quality estimation

The system analyzes image features to determine the level of handwriting quality. This process considers factors such as character sharpness, background noise level, blur, and geometric distortion. The output is a quality label corresponding to each data level.

4.1.2. Selection of the appropriate recognition model

Based on the identified quality label, the system selects a suitable model according to the experimental findings:

- **High quality (printed text):** lightweight models such as CNN or ResNet are used to ensure high processing speed.
- **Medium quality (clear handwriting):** the ResNet + BiLSTM + CTC model is applied to effectively capture spatial features and character sequence dependencies.
- **Low quality (poor handwriting):** Transformer-based models such as TrOCR are used to improve contextual understanding and robustness against noise.

4.2. Practical Application Process of the Framework

In real-world text recognition systems, the proposed framework acts as a coordination layer between input data and processing models. The workflow consists of the following steps:

- **Step 1: Input image acquisition:** the system receives the document image to be recognized.
- **Step 2: Text quality analysis:** the system determines the image quality level based on visual characteristics.
- **Step 3 Model selection:** an appropriate model is chosen according to the predefined mapping rules.
- **Step 4: Text recognition:** the selected model performs inference and outputs the recognized text.

This approach enables the system to adapt flexibly to heterogeneous data instead of relying on a single fixed model.

4.3. Significance of the Proposed Framework

The proposed text quality-based model selection framework offers several practical advantages:

- First, it improves overall performance by selecting models that are better suited to different types of data.
- Second, it reduces computational cost by avoiding the use of complex models for simple inputs.
- Third, it enhances system adaptability in real-world environments where data is diverse and non-uniform.
- Finally, it provides a systematic decision-making mechanism instead of relying on empirical model selection.

5. Experimental Validation of The Model Selection Framework

After constructing the model selection framework based on handwriting quality levels in Section 4, experiments were conducted to evaluate the stability and applicability of the proposed framework under different data conditions.

Unlike the initial experiments, this section does not focus on comparing models, but instead aims to verify the effectiveness of the proposed framework when applied to new datasets, thereby validating the correctness of the quality-based model selection strategy.

In the validation stage, three independent datasets different from those used in the initial experiments are employed, corresponding to three handwriting quality levels: the printed text level uses the ICDAR2013 Scene Text Dataset [14], which contains clear text images in natural scenes; the clear handwriting level uses the CVL Database [15] with an independent test split; and the degraded handwriting level uses data collected from students, along with data augmentation techniques such as blurring, Gaussian noise injection, and geometric distortions to simulate complex real-world conditions.

All datasets are strictly separated into training and testing sets to ensure objectivity in evaluation.

5.1. Experimental Setup

In the validation experiments, the models are trained on a synthetic dataset similar to that described in Section 3, and the model selection framework proposed in Section 4 is then applied during inference.

Specifically, for each input image, the system first determines the handwriting quality level. Based on this classification result, an appropriate model is automatically selected according to the proposed framework. The recognition outputs are then collected for performance evaluation.

The models used in the experiments include CNN (baseline), ResNet + BiLSTM + CTC, and TrOCR and the evaluation metrics include Accuracy, CER, and WER.

5.2. Experimental Results

Table 2: Validation results of the proposed framework

Data level	Model selection strategy	Accuracy (%)	CER (%)	WER (%)
ICDAR2013 Scene Text Dataset	CNN (baseline)	0.90%	0.13%	0.20%
CVL Database	ResNet + BiLSTM + CTC	0.86%	0.18%	0.27%
Poor handwriting	TrOCR	0.84%	0.22%	0.31%

5.3. Analysis of Results

The experimental results show that the proposed model selection framework performs stably across all three data levels.

At the printed text level, the lightweight CNN (baseline) still achieves high performance, indicating that using more complex models is unnecessary for simple data conditions.

At the clear handwriting level, the ResNet + BiLSTM + CTC model provides a balanced performance in terms of accuracy and sequential text modeling capability, making it suitable for intermediate data characteristics.

Meanwhile, at the degraded handwriting level, TrOCR achieves the best performance due to its ability to model global context and effectively handle noise.

Notably, the results indicate that when the appropriate model is selected according to the proposed framework, the overall system performance remains stable, without significant degradation as observed when using a single fixed model for all data types.

5.4. Comparison Chart

The chart below illustrates the Accuracy performance of the system when applying the model selection framework across different handwriting quality levels.

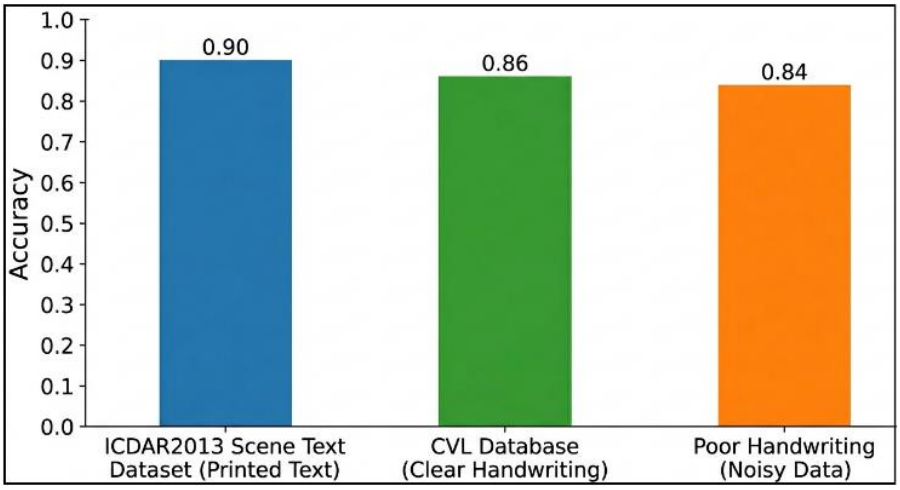

Fig 2: Model Performance Across Handwriting Levels.

The results show that the system maintains stable performance across all three data levels, and is particularly effective in mitigating significant performance degradation on low-quality data, thereby demonstrating the effectiveness of the proposed framework.

The validation results confirm that selecting models based on handwriting quality yields clear improvements compared to using a single fixed model approach.

The proposed framework not only enhances performance at each data level but also improves the system's adaptability to real-world, heterogeneous data.

However, one limitation of the current approach is that image quality estimation still relies on handcrafted features or simple assumptions. In future work, deep learning-based models could be integrated to automatically classify image quality, thereby further improving the system's level of automation.

6. Discussion

6.1. Significance of the Experimental Results

The experimental results show that the performance of text recognition models is strongly dependent on the quality level of the input handwriting. At the same time, applying a quality-based model selection framework improves the overall stability of the system under different data conditions.

Instead of relying on a single model to handle all inputs, the system dynamically selects the most suitable model for each data quality level. This helps maintain stable performance across printed text, clear handwriting, and degraded handwriting, while reducing performance degradation on difficult samples.

The results also confirm that the alignment between data characteristics and model architecture is a key factor in real-world text recognition tasks.

6.2. Practical Impact

The proposed framework has clear practical significance in document processing and text recognition systems.

In applications such as document digitization, the system can automatically select appropriate models to optimize both speed and accuracy. For clean inputs such as printed text, lightweight models can be used to reduce computational cost.

In contrast, for more complex inputs such as handwriting or degraded documents, stronger models can be applied to ensure higher accuracy.

This improves real-world deployability, especially in multi-source document systems where input data is heterogeneous and continuously changing.

6.3. Comparison with Traditional Evaluation Methods

Traditional approaches in text recognition typically evaluate models using a single mixed test set and report overall average performance. Although simple, this approach does not fully reflect the diversity of data quality in real-world scenarios.

In contrast, the proposed framework separates data based on handwriting quality levels and evaluates each case independently. This provides a clearer understanding of model behavior under different conditions, rather than only reporting a single aggregated score.

Moreover, traditional methods do not support model selection during deployment, whereas the proposed framework can act as a decision layer that automatically selects the most suitable model for each input.

6.4. Limitations and Future Work

Despite its effectiveness, the proposed framework still has several limitations.

First, the current handwriting quality estimation is based on qualitative criteria or dataset assumptions, and does not yet include a fully automated or precise quantitative mechanism.

Second, the framework does not fully consider document structural information such as page layout, relationships between text regions, or contextual document structure, which may affect performance in complex documents.

Third, the current approach is static and does not support adaptive or dynamic model selection based on runtime feedback or user-specific data.

In future work, the following directions can be explored:

- Integrating deep learning models for automatic input quality assessment.
- Incorporating document structural features such as layout or graph-based representations.
- Developing dynamic model selection mechanisms based on performance feedback over time.

7. Conclusion

In this study, we propose a handwriting quality-based model selection framework to improve the performance of text recognition systems in document images. The framework is built on the experimental observation that model performance varies significantly across different data types, including printed text, clear handwriting, and degraded handwriting, enabling appropriate model selection instead of relying on a single fixed model.

Experimental results with CNN, ResNet + BiLSTM + CTC, and TrOCR show that the proposed framework improves system stability and better adapts to different data quality levels. However, the current approach still has limitations in automatically estimating handwriting quality and in leveraging document structural information, which will be addressed in future work.

References

1. Plamondon R, Srihari SN. Online and off-line handwriting recognition: a comprehensive survey. IEEE Trans Pattern Anal Mach Intell. 2000;22(1):63–84. doi:10.1109/34.824821
2. Szeliski R. Computer Vision: Algorithms and Applications. Springer; 2010. Available from: https://szeliski.org/Book/
3. Chen K, Seuret M. Convolutional neural networks for page segmentation of historical document images. arXiv preprint arXiv:1704.01474; 2017. Available from: https://arxiv.org/abs/1704.0147
4. Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. In: Proc IEEE Conf Comput Vis Pattern Recognit (CVPR). 2016. Available from: https://arxiv.org/abs/1507.05717
5. Fujitake M. DTrOCR: Decoder-only transformer for optical character recognition. arXiv preprint arXiv:2308.15996; 2023.
6. Li C, *et al*. TrOCR: Transformer-based optical character recognition with pre-trained models. 2021. Available from: https://arxiv.org/abs/2109.10282
7. Li C, *et al*. TrOCR: Transformer-based optical character recognition with pre-trained models. 2021. Available from: https://arxiv.org/abs/2109.10282
8. Bautista M, *et al*. PARSeq: Scene text recognition with permuted autoregressive sequence models. 2022. Available from: https://arxiv.org/abs/2207.06966
9. Mishra A, Alahari K, Jawahar CV. IIIT 5K-Words dataset for scene text recognition. CVIT, International Institute of Information Technology Hyderabad. Available from: https://cvit.iiit.ac.in/research/projects/cvit-projects/the-iiit-5k-word-dataset
10. Marti UV, Bunke H. IAM handwriting database: an English sentence database for offline handwriting recognition. University of Bern, Institute of Computer Science and Applied Mathematics. Available from: https://fki.tic.heia-fr.ch/databases/iam-handwriting-database
11. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2:37–63.
12. Neudecker C, *et al*. A survey of OCR evaluation tools and metrics. In: Proc ACM Symp Document Engineering (DocEng). 2021. doi:10.1145/3476887.3476888
13. Kolak O, Resnik P. OCR error correction using a noisy channel model. In: Proc HLT-NAACL. 2003.
14. ICDAR 2013 robust reading competition: scene text detection and recognition dataset. CVL, CVC Barcelona. Available from: https://rrc.cvc.uab.es/?ch=2
15. Fischer M, Wimmer M, Bunke H, Kaufmann G. A comprehensive offline handwriting database for writer retrieval, writer identification and word spotting (CVL database). TU Wien, Computer Vision Lab. Available from: https://cvl.tuwien.ac.at/research/cvl-databases/an-off-line-database-for-writer-retrieval-writer-identification-and-word-spotting/

How to Cite This Article

Nguyen Thi Thanh P, Pham Phu K, Tran Van T, Ngo Van H. A model selection framework based on writing quality levels for image-based text recognition. International Journal of Future Engineering Innovations. 2026;3(3):20-25. doi:10.54660/IJFEI.2026.3.3.20-25.