JTE HCM-UTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Engineering
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Automatic Information Extraction From Student ID Card Images Using DB and VietOCR: A Case Study at a Vietnamese University

Bao-Khanh Hoang[1], Van-Hai Ngo[1], Xuan-Truong Tran[1], Dung Nguyen[1*], Duc-Phuc Nguyen[2]
[1]*University of Sciences, Hue University, Vietnam*
[2]*University of Arts, Hue University, Vietnam*

[*]Corresponding author. *Email: nguyendung@hueuni.edu.vn*

ARTICLE INFO

Received: 11/02/2026
Revised: 23/02/2026
Accepted: 06/04/2026
Online First: 29/04/2026
Published:

KEYWORDS

MobileNetV3;
Differentiable Binarization;
Text Detection;
Vietnamese Text Recognition;
VietOCR.

ABSTRACT

The development of an information extraction system from student ID card images plays an important role in the digitalization of student management. This study proposes a two-stage processing framework that integrates computer vision and deep learning techniques, in which MobileNetV3-Small is employed for student identification card image classification, while the Differentiable Binarization (DB) model and VietOCR are responsible for Vietnamese text detection and recognition, respectively. Experimental results on a student ID card image dataset show that the classification model achieves an accuracy of 99.40% with an AUC of 0.9996, while the DB-based text detection model attains an Hmean of 89.81% after data augmentation. For text recognition, the proposed system achieves over 99% character-level accuracy and up to 98.90% full-sequence accuracy. These results demonstrate the effectiveness and practical feasibility of the proposed system, which is further validated through a proof-of-concept offline attendance application. In addition, the system is designed with computational efficiency in mind, enabling deployment on resource-constrained devices without requiring continuous internet connectivity. The proposed framework can be readily adapted to other types of identification documents, providing a scalable and cost-effective solution for automated data acquisition in educational institutions.

Doi: https://doi.org/10.54644/jte.2026.2101

1. Introduction

In the context of the rapid digital transformation taking place in educational institutions, the management and digitization of information from paper-based documents, particularly student identification (ID) cards, has become an urgent requirement [1]–[3]. Student ID cards not only serve as a means of personal identification in academic and administrative activities but also contain various important data fields such as full name, student ID number, date of birth, and academic information. However, traditional data extraction methods relying on manual operations reveal several limitations, including high time and labor costs, susceptibility to human errors, difficulties in ensuring data consistency, and limited scalability as data volumes continue to grow [4]. These issues hinder the modernization of information management systems in universities and colleges.

Several technological solutions have been studied and applied to automate this process, notably the integration of Near Field Communication (NFC) or Radio-Frequency Identification (RFID) chips into student ID cards [5], [6]. Nevertheless, these approaches still suffer from significant drawbacks: (1) they require substantial investment costs to replace existing card systems [7], (2) device compatibility is inconsistent, as NFC functionality is typically supported only on mid-range and high-end devices, and (3) security risks arise because RFID data can be easily accessed without robust encryption mechanisms [8], [9]. This gap highlights the need for an alternative approach that is both effective and cost-efficient, while also ensuring flexible deployment on existing technological infrastructures.

Based on these considerations, this study proposes an information extraction system for student ID card images that integrates three main components: (1) MobileNetV3-Small [10] for student ID card

image classification; (2) Differentiable Binarization (DB) [11] for text region detection; and (3) VietOCR [12] for Vietnamese character recognition. This combination aims to develop an automated solution with high accuracy, optimized cost, seamless integration into existing data management systems, and strong suitability for the characteristics of the Vietnamese language.

2. Related Work

2.1. Document Information Extraction Systems

Automated information extraction from identity documents has transitioned from rule-based methods to deep learning approaches over the past decade. Early systems relied heavily on Machine Readable Zones (MRZ) and predefined templates [13], which proved inadequate for modern identity cards that lack standardized formats and exhibit significant layout variability across different institutions and countries.

Recent studies demonstrate that deep learning-based approaches, particularly those combining instance segmentation with Optical Character Recognition (OCR), significantly outperform traditional methods when handling diverse document formats [4]. However, a large portion of existing work primarily focuses on Western languages with relatively simple character systems. In contrast, Vietnamese text, characterized by complex diacritics and multiple tonal marks, poses unique challenges that remain underexplored in the literature.

While commercial OCR systems such as Google Vision API and Tesseract achieve high accuracy on English documents [14], their performance degrades substantially on Vietnamese text, particularly for student ID cards with varied fonts, small text sizes, and complex background textures. This limitation highlights the need for domain-specific OCR solutions tailored to Vietnamese document characteristics.

Recent transformer-based Document AI models further demonstrate the importance of jointly modeling visual layout and textual content. LayoutLMv3 [15] introduces unified masking objectives over both text and image tokens for document understanding, while Donut [16] proposes an OCR-free end-to-end transformer that directly maps document images to structured outputs. Although these methods report strong results on benchmark datasets, they are generally designed for broader document understanding tasks and may require substantial pre-training resources, motivating more lightweight pipelines for constrained deployment scenarios.

2.2. Lightweight Neural Network Architectures for Edge Deployment

The deployment of computer vision systems on resource-constrained devices necessitates a careful balance between model accuracy and computational efficiency. Lightweight convolutional neural networks have therefore attracted significant attention.

The MobileNet series pioneered the use of depthwise separable convolutions to reduce computational cost while maintaining competitive accuracy [17]. Building upon this foundation, Howard et al. introduced MobileNetV3 [10], which integrates hardware-aware Neural Architecture Search (NAS) with the NetAdapt algorithm. MobileNetV3-Large improves ImageNet classification accuracy by over 3% while reducing latency by approximately 20% compared to MobileNetV2, whereas MobileNetV3-Small offers even greater efficiency for low-resource scenarios.

Although MobileNetV3 demonstrates strong performance on generic image classification benchmarks, its effectiveness for document analysis tasks, which require fine-grained text localization rather than global semantic understanding, remains less explored. Furthermore, many reported benchmarks rely on high-end mobile processors, while real-world deployments such as student ID scanning systems often operate on low-cost smartphones or embedded devices. In this context, MobileNetV3-Small provides an attractive trade-off between accuracy and efficiency, particularly when combined with task-specific fine-tuning on document images.

2.3. Scene Text Detection Methods

Scene text detection has evolved from traditional sliding-window techniques to modern segmentation-based approaches capable of handling text with arbitrary shapes and orientations. A key

advancement in this area is DB [11]. DB addresses a major limitation of previous segmentation-based methods, namely the use of fixed thresholding during post-processing. By introducing a trainable binarization module that approximates the step function with a differentiable formulation, DB enables end-to-end optimization of the entire text detection pipeline.

With a ResNet-18 backbone, DB achieves an F-measure of 82.8% at 62 FPS on the MSRA-TD500 dataset, and further improves accuracy when using deeper backbones. Compared to alternative approaches such as EAST [18], TextSnake [19], or PSENet, DB offers a favorable balance of detection accuracy, computational efficiency, and implementation simplicity. These properties make DB particularly suitable for Vietnamese student ID cards, as further confirmed by the experimental results presented in Section 4.

2.4. Attention Mechanisms for Sequence Recognition

Optical Character Recognition has progressed from convolution-only architectures to attention-based encoder–decoder models that better capture sequential dependencies in text. The attention mechanism was first introduced in [20] to address the limitations of encoding variable-length input sequences into fixed-length representations. By allowing the decoder to dynamically attend to relevant parts of the encoder output at each time step, attention-based models significantly improve recognition performance on long and complex sequences.

For Vietnamese text recognition, attention mechanisms offer several advantages, including improved handling of diacritical marks, natural support for variable-length outputs, and enhanced contextual modeling to disambiguate visually similar characters. However, attention-based models generally require larger amounts of annotated data compared to CTC-based approaches [21]. This limitation motivates the use of transfer learning and careful fine-tuning strategies when working with limited Vietnamese document datasets.

In addition, transformer-based OCR models such as TrOCR [22] have shown that pre-trained vision–language architectures can deliver highly competitive recognition accuracy across both printed and handwritten text. This trend confirms the effectiveness of transfer learning for OCR, but also highlights the computational trade-offs that must be considered for real-time and edge-oriented applications.

2.5. Transfer Learning and Pre-trained Models

Transfer learning has become indispensable for document analysis tasks with limited annotated data. The ImageNet dataset [23] provides large-scale supervision that enables models to learn generic visual representations transferable to downstream tasks.

Prior studies have shown that ImageNet-pretrained backbones can improve text detection Intersection over Union (IoU) by 15–20% compared to random initialization, and achieve over 90% accuracy on document analysis tasks with relatively small training datasets [24]. Nevertheless, a domain gap exists between natural image classification and document image understanding, necessitating additional fine-tuning on document-specific data. Motivated by these observations, our work leverages ImageNet-pretrained models combined with task-specific fine-tuning on Vietnamese student ID card images [25].

The reviewed literature reveals several under-addressed challenges: (1) limited research on Vietnamese text with complex diacritics, (2) unrealistic deployment assumptions that overlook resource-constrained devices, (3) scarcity of annotated Vietnamese student ID datasets, and (4) a lack of end-to-end system evaluations that consider error propagation across classification, detection, and recognition stages. To address these gaps, this work proposes a complete Vietnamese student ID information extraction pipeline integrating MobileNetV3-based card classification, DB-based text detection, and attention-based text recognition. Unlike prior studies that focus on individual components in isolation, we provide a holistic evaluation of accuracy, efficiency, and real-world deployment constraints on Vietnamese student ID card data.

3. Methodology

The student ID card recognition and classification system proposed in this study is designed as a pipeline consisting of three main stages: card classification, text detection, and information recognition from the card. The detailed architecture of the system is illustrated in Figure 1.

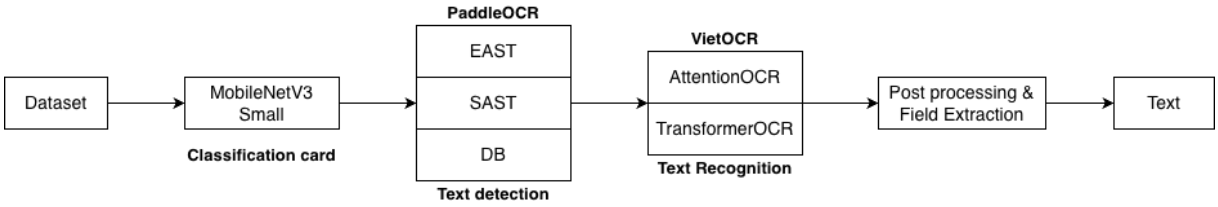

Figure 1. *Overall Pipeline for Information Extraction from Student ID Card Images*.

3.1. Card Classification Using MobileNetV3-Small

In the first stage of the proposed system, the card image classification task is essential for determining whether an input image corresponds to a ID card or a library card. An appropriate model for this task must satisfy two main requirements: (1) high classification accuracy to minimize errors in subsequent information extraction stages, and (2) the ability to be deployed on resource-constrained devices in practical settings [17], [10]. We evaluated several models from the MobileNet family and selected MobileNetV3-Small as the backbone for classification. MobileNetV3-Small [10] is an improved variant of its predecessors, MobileNet and MobileNetV2 [17], featuring hardware-aware neural architecture search and optimized layers for efficiency.

For this study, input card images are preprocessed by resizing them to 224×224 pixels and normalizing pixel values to the range $[0, 1]$. The MobileNetV3-Small model is initially pre-trained on the ImageNet1K dataset [23] and subsequently fine-tuned using the student ID and library card image datasets. The original fully connected layer is replaced with a custom classification head structured as follows:

Global Average Pooling 2D → Dense (ReLU) → Dropout → Dense (Softmax)

During fine-tuning, the backbone weights are frozen due to the relatively small and domain-specific dataset, which helps mitigate the risk of overfitting.

3.2. Text Detection Using Differentiable Binarization

Differentiable Binarization (DB) [11] is a trainable binarization technique that enables end-to-end optimization of the entire pipeline. By eliminating the need for manually defined thresholding, DB improves both inference speed and detection accuracy. The detailed architecture of DB is illustrated in Figure 2.

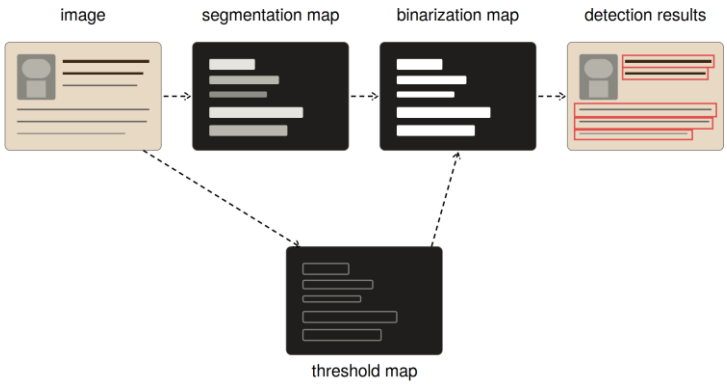

Figure 2. *Application of the DB model on the student ID card image dataset*.

Overall, the DB model [11] is a simple yet effective segmentation-based network for scene text detection. The input to the network is the original image, which is processed by a CNN backbone (e.g., ResNet-18/50 or MobileNetV3). The extracted features are fused to produce a shared feature map F, from which the network branches into two main outputs: a probability map P and a threshold map T. The probability map P represents the likelihood of each pixel belonging to a text region, while the threshold map T provides an adaptive binarization threshold for each pixel.

Given P and T, a differentiable binarization function is applied to generate an approximate binarization map B. During training, all three maps P, T, and B are supervised (with P and B sharing the same ground-truth labels). During inference, the final detection results are obtained from either the binarization map or the probability map through a box formulation module that generates bounding boxes for text regions.

3.3. Text Recognition Using VietOCR

VietOCR [12] is an open-source library for Optical Character Recognition (OCR), specifically designed and optimized for the Vietnamese language. The library supports input data in the form of images or scanned documents and outputs machine-readable text that can be searched and edited. VietOCR provides two main OCR models, namely AttentionOCR and TransformerOCR [12]. After evaluating both models, we selected AttentionOCR as the text recognition model for this study.

In AttentionOCR, the Sequence-to-Sequence (Seq2Seq) model [26] is employed to address tasks that map an input sequence of variable length to a corresponding output sequence, such as machine translation, as illustrated in Figure 3. The Seq2Seq architecture consists of two main components: an Encoder and a Decoder. The Encoder maps the input sequence into a set of hidden states and compresses the encoded information into a context vector. The Decoder is then initialized from this context vector and generates the elements of the output sequence in a sequential manner [20].

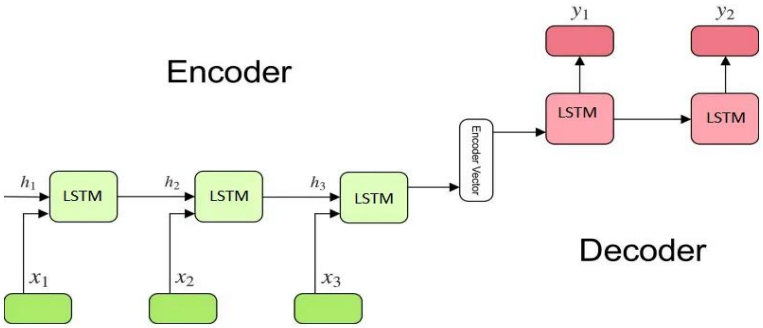

Figure 3. *Seq2Seq model using LSTM [27].*

The attention mechanism was introduced in 2015 [20] to address the context degradation problem in Seq2Seq models when processing long character sequences. The core idea of attention is that not all input tokens contribute equally to the generation of an output token. The attention mechanism enables the decoder to directly attend to all hidden states of the encoder at each time step. Instead of relying on a fixed context vector, attention computes a dynamic context vector by assigning attention weights to each encoder hidden state, reflecting their relevance to the token being generated [28]. Consequently, the Seq2Seq model augmented with the attention mechanism forms the foundation of the widely adopted AttentionOCR model, as illustrated in Figure 4.

In deep learning models, the stride size refers to the step size of the convolution kernel or pooling window as it slides over an image or feature map. To enhance model performance in the context of student ID card image data with asymmetric aspect ratios, we adjust the stride size in the pooling layers of VGG19 [30]. Instead of using the default stride of $[2, 2]$, we apply a stride of $[2, 1]$ in the later pooling layers, reducing the spatial resolution only along the height dimension while preserving the width. This adjustment helps retain horizontal character information and significantly improves recognition accuracy.

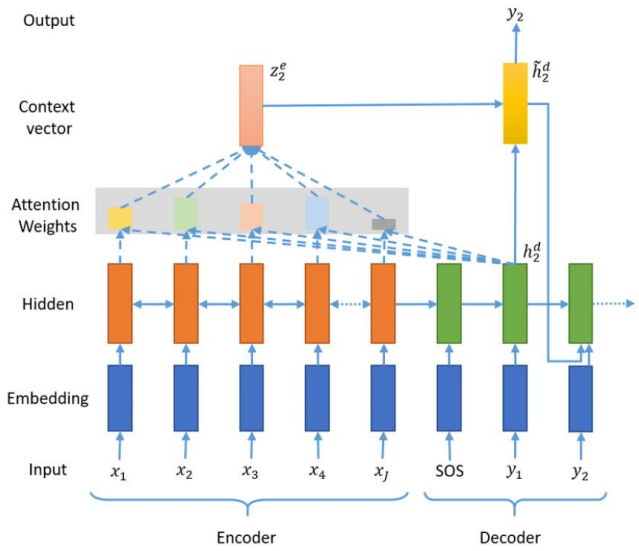

Figure 4. *Seq2Seq model with an attention mechanism [29].*

4. Experimental Results and Discussion

4.1. Dataset

The proposed system was trained and evaluated on a student ID card image dataset collected at the University of Sciences, Hue University. The dataset consists of 555 images, including 118 student ID card images and 437 library card images. All images were captured using smartphones under standard lighting conditions with a frontal viewpoint. This dataset reflects real-world usage scenarios and exhibits moderate image quality, making it a valuable resource for training and evaluating the models within the proposed system. Figure 5 illustrates the two types of cards included in the dataset.

Figure 5. *Two types of cards in the student ID card image dataset (Student ID card on the left, library card on the right. Images have been blurred to protect personal information.).*

The two card types in the dataset are separated into two individual subsets, namely the student ID card set and the library card set. Each subset is further divided into training and evaluation sets with an 8:2 ratio. The dataset is utilized to conduct three main experiments:

- **Card classification training and evaluation:** In this experiment, the training sets of both card types are merged to form a unified training set, and the evaluation sets are combined in the same manner.
- **Text detection training and evaluation:** Data augmentation is applied to the training set to alleviate the data imbalance between the student ID card set and the library card set. Specifically, the student ID card images are augmented to achieve a quantity approximately equal to that of the library card images.
- **Text recognition training and evaluation:** Text lines in each image, including full name, date of birth, class, academic year, and student ID number, are cropped and used for training the text recognition model.

Although the collected dataset reflects realistic acquisition conditions, it remains relatively small, consisting of only 555 images, including 118 student ID cards and 437 library cards. The limited number of student ID samples may restrict the diversity of layout variations, font styles, printing quality, and capture conditions represented in the training data. Furthermore, since all images were collected from a single university, the dataset may not fully capture inter-institutional variability in ID card design across different universities in Vietnam. This could limit the generalization capability of the trained models when deployed in broader contexts. Although data augmentation techniques were applied to mitigate class imbalance and improve robustness, synthetic transformations cannot fully substitute real-world diversity. Therefore, future work should focus on constructing a larger multi-institutional dataset to enhance the scalability and generalization performance of the proposed system.

4.2. Implementation Details

To improve reproducibility, we summarize the key implementation settings used in our experiments.

- **Hardware and software environment.** Training and inference were conducted on a workstation with a single NVIDIA GPU using the PyTorch framework.
- **Input configuration.** For card classification, input images were resized to 224×224 and normalized to $[0, 1]$. For text detection, full card images were resized while preserving aspect ratio and padded to a fixed input size for mini-batch training. For text recognition, cropped text lines were resized to a fixed height while preserving width-related character structure.
- **Optimization settings.** All models were fine-tuned from pre-trained weights using the Adam optimizer with mini-batch training and early stopping based on validation performance. The initial learning rate was set in the range 10^{-4} to 10^{-3} depending on the module, and reduced using a scheduler when validation metrics plateaued.
- **Data split and evaluation protocol.** The card classification module followed 5-fold cross-validation on the training partition, while text detection and text recognition used fixed train/validation splits derived from the 8:2 data partition described above.

These settings, together with the reported model architectures and evaluation metrics, improve the reproducibility of the proposed experimental pipeline.

4.3. Evaluation Methodology

4.3.1. Card Classification Model

Common metrics for binary classification tasks, including Accuracy, Precision, Recall, F1-score, and Area Under the Curve (AUC), are employed to evaluate the performance and reliability of the card classification model. Among these, Precision, Recall, and F1-score are computed using both weighted and macro averaging schemes. This dual evaluation strategy aims to reflect the model performance under practical conditions, such as class imbalance, as well as under ideal conditions where all classes are assigned equal importance.

4.3.2. Text Detection Model

The text detection model is evaluated based on two main criteria: detection quality and practical deployment performance. The former measures the accuracy and reliability of the model, while the latter assesses its operational efficiency on resource-constrained devices and under real-time processing requirements. Detection quality:

- Intersection over Union (IoU): Measures the overlap between the predicted bounding box and the ground-truth bounding box. A higher IoU indicates a closer match between the predicted and ground-truth bounding boxes.
- Precision, Recall, and F1-score (Hmean): Evaluate the detection effectiveness of the model. The values of True Positives (TP), False Positives (FP), and False Negatives (FN) are determined based on the IoU metric and a predefined threshold (typically 0.5).
- Inference Time: Measures the processing time required for a single data sample. A shorter inference time indicates higher computational efficiency of the model.

4.3.3. Text Recognition Model

During the evaluation of the text recognition model, Precision Full Sequence and Precision Per Character are employed to assess the accuracy and reliability of the recognition results. These two metrics measure performance at different levels: sequence-level accuracy and character-level accuracy.

- Precision Full Sequence: Measures accuracy at the full-sequence level. A predicted sequence is considered correct only if it exactly matches the corresponding ground-truth sequence.
- Precision Per Character: Measures accuracy at the character level, allowing a more fine-grained evaluation of the model's ability to recognize individual characters.

In addition to recognition accuracy, Inference Time is also used to evaluate the processing efficiency of the text recognition model.

4.4. Experimental Results

4.4.1. Evaluation of the Card Classification Model

We conducted a comparative study using models from the MobileNet family, including MobileNet, MobileNetV2, MobileNetV3-Small, and MobileNetV3-Large, as backbones for the card classification task. All backbone networks were pre-trained on the ImageNet-1K dataset [20]. To address the limited size of the dataset, a k-fold cross-validation strategy with $k = 5$ was applied on the training set. The final performance was obtained by aggregating the evaluation results across the five folds. Since the dataset exhibits class imbalance, class-weighting was employed during training to mitigate its negative impact on model performance. The weight assigned to each class is computed as follows in Eq. (1):

$$\text{weight}_y = \frac{n_{\text{sample}}}{n_{\text{classes}} \times \text{freq}_y}, \tag{1}$$

where weight_y denotes the weight of class y, n_{sample} is the total number of samples, n_{classes} is the number of classes, and freq_y represents the number of samples belonging to class y. Table 1 presents the comparative results of different backbone networks on the card classification task.

Table 1. *Comparison of classification performance using different backbone networks. P, R, and F1 denote Precision, Recall, and F1-score.*

Backbone	Accuracy	Macro			Weighted			AUC
		P	R	F1	P	R	F1	
MobileNet	98.81	97.37	99.24	98.27	98.87	98.81	98.82	0.9989
MobileNetV2	98.93	97.68	99.32	98.45	99.00	98.93	98.94	0.9995
MobileNetV3-Small	99.40	98.65	99.62	99.12	99.42	99.40	99.41	0.9996
MobileNetV3-Large	99.40	98.65	99.62	99.12	99.42	99.40	99.41	0.9990

Based on the results in Table 1, both MobileNetV3-Small and MobileNetV3-Large achieve the best and comparable performance across all evaluation metrics. However, MobileNetV3-Small attains a slightly higher AUC score than MobileNetV3-Large, indicating better generalization capability. Considering its superior performance and lower computational complexity, MobileNetV3-Small is selected as the backbone for the card classification module in the proposed student ID information extraction system.

4.4.2. Evaluation of the Text Detection Model

To evaluate the effectiveness of text detection models, we conducted experiments on two widely used approaches, namely DB-MV3 and EAST-MV3, under two different settings: without data augmentation and with data augmentation. The evaluation metrics include Precision (P), Recall (R), and Harmonic Mean (Hmean), which are commonly adopted in text detection tasks. Table 2 presents the comparative experimental results of DB-MV3 and EAST-MV3 under the two data settings.

Table 2. *Comparison of text detection performance between DB-MV3 and EAST-MV3. P, R, and H denote Precision, Recall, and Hmean.*

Model	Without data augmentation			With data augmentation		
	P	R	H	P	R	H
DB-MV3	85.14	84.00	84.56	87.26	92.50	89.81
EAST-MV3	56.13	74.83	64.14	68.04	77.00	72.24

The experimental results demonstrate that DB-MV3 consistently outperforms EAST-MV3 across all evaluation metrics. In the setting without data augmentation, DB-MV3 achieves an Hmean of 84.56, which is significantly higher than that of EAST-MV3 (64.14), indicating superior overall text detection capability. After applying data augmentation techniques, the performance of both models improves noticeably. DB-MV3 continues to exhibit outstanding performance, achieving an Hmean of 89.81. Moreover, its Precision and Recall increase to 87.26 and 92.50, respectively, confirming that the model is able to detect more text regions while reducing detection errors. These results validate the robustness and effectiveness of the DB-MV3 model for text detection in student ID card images.

4.4.3. Evaluation of the Text Recognition Model

During the training and evaluation phases, we investigated two backbone architectures corresponding to the AttentionOCR and TransformerOCR models, namely vgg_seq2seq and vgg_transformer. The performance was evaluated using two metrics: Precision Full Sequence and Precision Per Character, which measure recognition accuracy at the sequence level and character level, respectively. Table 3 presents the experimental results obtained on the two backbones under two different stride size configurations: the default stride size of [2, 2] and the adjusted stride size of [2, 1].

Table 3. *Text recognition performance on vgg_seq2seq and vgg_transformer backbones.*

Backbone	Default stride [2,2]		Adjusted stride [2,1]	
	Full Sequence	Per Char	Full Sequence	Per Char
vgg_seq2seq	98.90	99.49	98.76	99.35
vgg_transformer	96.80	99.43	97.45	99.68

The experimental results show that vgg_seq2seq with adjusted stride achieves the highest Full Sequence Accuracy (98.76%), while vgg_transformer attains slightly higher character-level accuracy (99.68%). Considering the trade-off between sequence-level reliability and computational efficiency, vgg_seq2seq was selected for the proposed system. Furthermore, the Per Character Precision exceeds 99% for both backbones, demonstrating the high character-level recognition capability of VietOCR. Despite its superior accuracy, especially on longer sequences, the vgg_transformer model requires significantly higher computational cost and memory resources. Transformer-based architectures are primarily designed for large-scale natural language processing tasks and typically rely on extensive training data, which does not fully align with the scale and practical requirements of the current project. Therefore, considering the trade-off between recognition accuracy, computational efficiency, and deployment feasibility, the vgg_seq2seq backbone of the AttentionOCR model is selected for the text recognition component of the proposed student ID card information extraction system.

4.4.4. Inference Time on CPU Platform

To evaluate practical feasibility without GPU acceleration, we measured inference time on an Intel Core i9-12900K CPU with 32GB RAM. The average processing time per image is summarized as follows:

- Card classification (MobileNetV3-Small): ~6–8 ms
- Text detection (DB-MV3): ~85–110 ms
- Text recognition (AttentionOCR – vgg_seq2seq): ~35–50 ms

- Total pipeline inference time: ~130–160 ms per image

The results indicate that the text detection module accounts for the largest computational cost, while classification contributes only minimal overhead. The total processing time corresponds to approximately 6–8 images per second on CPU, which is sufficient for offline attendance applications.

For comparison, preliminary testing shows that the vgg_transformer backbone requires approximately $1.5 - 2\times$ longer inference time than vgg_seq2seq on CPU due to the quadratic complexity of the self-attention mechanism. Therefore, despite slightly competitive character-level accuracy, vgg_seq2seq was selected as a better trade-off between accuracy and computational efficiency for practical deployment.

4.4.5. Error Propagation Analysis in the Multi-Stage Pipeline

Since the proposed system follows a sequential pipeline architecture, errors occurring in earlier stages may propagate and negatively affect downstream components.

First, misclassification at the card classification stage (e.g., an ID card classified as a library card) prevents subsequent text detection and recognition, resulting in complete information extraction failure. Although the classification accuracy reaches 99.40%, this stage remains a critical gatekeeper of the entire system. Second, errors in text detection (false negatives or inaccurate bounding boxes) directly impact the recognition stage. Missed text regions lead to incomplete information extraction, while imprecise bounding boxes may introduce background noise that degrades OCR accuracy. The detection Hmean of 89.81% therefore effectively constrains the upper bound of the end-to-end system performance. Third, although the character-level recognition accuracy exceeds 99%, minor errors may still lead to invalid administrative records, particularly for critical fields such as Student ID numbers and full names. A qualitative analysis of misrecognized samples shows that most errors arise from (1) visually similar characters such as “0/O” and “1/I”, (2) incorrect prediction of Vietnamese diacritics due to their small size and sensitivity to contrast variations, and (3) environmental factors such as mild glare, uneven illumination, or imperfect detection bounding boxes. In long text sequences, even a single character error results in a full-sequence mismatch, explaining the gap between character-level and sequence-level accuracy.

Overall, the detection stage contributes most significantly to end-to-end performance degradation, while recognition errors, though limited in number, have high practical impact in administrative scenarios. Future improvements should therefore prioritize enhancing detection robustness, incorporating structured post-recognition validation (e.g., format constraints for Student ID numbers), and exploring lightweight end-to-end optimization strategies to reduce inter-stage error amplification.

4.5. Qualitative Analysis and Limitations

Qualitative observations. In addition to the quantitative metrics, qualitative inspection indicates that DB-MV3 generally produces tighter and more complete text regions than EAST-MV3, especially for dense Vietnamese text lines. In the recognition stage, the selected AttentionOCR model preserves Vietnamese diacritics in common fields (e.g., full name and class), which is consistent with the high character-level precision in Table 3.

Current limitations. Despite strong average performance, the system can still degrade under challenging imaging conditions, including motion blur, partial occlusion, non-uniform illumination, strong highlights, and perspective distortion. In addition, the modular pipeline is affected by error propagation: inaccurate detection outputs may reduce downstream OCR quality.

Potential extensions. A promising direction is to move from a strictly modular pipeline toward layout-aware and end-to-end document understanding approaches. Integrating layout-aware pre-training (e.g., LayoutLMv3 [15]) or OCR-free paradigms (e.g., Donut [16]) may reduce intermediate hand-crafted steps and improve robustness on complex layouts.

5. Conclusion

This study proposed an automated system for extracting information from student ID card images by integrating MobileNetV3-Small for card classification, a DB-based model for text detection, and

VietOCR for Vietnamese text recognition. Experimental results demonstrate that MobileNetV3-Small achieves an accuracy of 99.40% with an AUC of 0.9996, confirming its effectiveness and efficiency for lightweight card classification. For text detection, the DB-MV3 model consistently outperforms the EAST-MV3 baseline, achieving an Hmean of 89.81% after data augmentation, which highlights its robustness in handling diverse layouts and challenging imaging conditions. In the text recognition stage, AttentionOCR with the vgg_seq2seq backbone delivers strong performance, reaching up to 98.90% full-sequence accuracy and exceeding 99% at the character level.

Beyond quantitative performance, the proposed pipeline demonstrates favorable characteristics in terms of robustness, modularity, and deployment feasibility. The use of lightweight architectures enables efficient inference on resource-constrained devices, making the system well-suited for practical applications. Moreover, the modular design allows each component to be independently optimized or replaced, facilitating future system upgrades.

Overall, the proposed system confirms its applicability to real-world scenarios such as student information digitization, identity verification, and offline attendance systems. Future work will focus on extending the framework to support additional identity document types, improving robustness under extreme capture conditions, and exploring end-to-end learning strategies to further enhance system efficiency and accuracy.

Acknowledgments

This research is supported by Hue University of Sciences, Hue University.

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability Statement

The data for this study were collected by the authors.

REFERENCES

[1] E. Mukul and G. Büyüközkan, "Digital transformation in education: A systematic review of Education 4.0," *Technol. Forecast. Soc. Change*, vol. 194, Art. no. 122664, 2023.
[2] K. K. de S. Oliveira and R. A. C. De Souza, "Digital transformation towards Education 4.0," *Informatics in Education*, vol. 21, no. 2, pp. 283–309, 2022.
[3] A. A. Bilyalova, D. A. Salimova, and T. I. Zelenina, "Digital transformation in education," in *Proc. Int. Conf. Integrated Science*, 2019, pp. 265–276.
[4] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: A survey," *Int. J. Doc. Anal. Recognit.*, vol. 7, no. 2–3, pp. 84–104, 2005.
[5] A. T. I. Mazumdar, N. N. Islam, and M. S. Hossain, "NFC-based mobile application for student attendance in institution of higher learning," in *Proc. ICAEEE*, 2022, pp. 1–6.
[6] M. Kumar, P. K. Samota, and M. K. Sharma, "Class attendance management system using NFC mobile devices," *Intell. Autom. Soft Comput.*, vol. 23, no. 2, pp. 243–250, 2017.
[7] T. Karygiannis *et al.*, *Guidelines for Securing Radio Frequency Identification (RFID) Systems*, NIST Special Publication 800-98, 2007.
[8] C. Jin *et al*., "RFID technology, security vulnerabilities, and countermeasures," in *Cutting Edge Research Topics on Multiple Access Communications*. London, U.K.: IntechOpen, 2009.
[9] S. Kumar *et al*., "A comprehensive taxonomy of security and privacy issues in RFID," *Complex Intell. Syst.*, vol. 7, no. 4, pp. 1915–1943, 2021.
[10] A. Howard *et al*., "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1314–1324.
[11] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11474–11481.
[12] P. B. C. Quoc, "VietOCR – Nhận dạng tiếng Việt sử dụng mô hình Transformer và AttentionOCR," 2021. [Online]. Available: https://pbcquoc.github.io/vietocr/
[13] A. V. Gayer, Y. S. Chernyshova, and V. V. Arlazarov, "Recognition of machine-readable zone in identity documents: A review," *IEEE Access*, 2025.
[14] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. ICDAR*, 2007, pp. 629–633.
[15] Y. Xu *et al*., "LayoutLMv3: Pre-training for document AI with unified text and image masking," arXiv:2204.08387, 2022.
[16] G. Kim *et al*., "Donut: Document understanding transformer without OCR," in *Proc. ECCV*, 2022, pp. 1–19.
[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, 2018, pp. 4510–4520.
[18] X. Zhou *et al*., "EAST: An efficient and accurate scene text detector," in *Proc. CVPR*, 2017, pp. 2642–2651.
[19] S. Long *et al*., "TextSnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. ECCV*, 2018, pp. 20–36.
[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv:1409.0473, 2014.
[21] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, 2017.

[22] M. Li *et al*., "TrOCR: Transformer-based optical character recognition with pre-trained models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 11, 2023, pp. 13094–13102.
[23] J. Deng *et al*., "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.
[24] Y. Xu *et al*., "LayoutLM: Pre-training of text and layout for document image understanding," in *Proc. ACM SIGKDD*, 2020, pp. 1192–1200.
[25] K. Nguyen-Trong, "An end-to-end method to extract information from Vietnamese ID card images," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 3, 2022.
[26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.
[27] P. Dhote, "Seq2Seq Encoder–Decoder LSTM Model," Medium, 2020.
[28] A. Vaswani *et al*., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[29] Viblo Asia, "Seq2Seq with Attention," 2019. [Online]. Available: https://viblo.asia
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

Bao-Khanh Hoang was born on November 15, 2004, in Hue. He is currently a fourth-year student in Information Technology, majoring in Computer Science at the University of Sciences, Hue University. Research areas: Artificial Intelligence, Computer Vision.
Email: 22T1020637@husc.edu.vn. ORCID: https://orcid.org/0009-0002-0269-9612

Van-Hai Ngo was born on September 9, 2003, in Hue. In 2026, he graduated with a bachelor's degree in Computer Science from the University of Sciences, Hue University. Research areas: Artificial Intelligence, Computer Vision.
Email: 21T1020340@husc.edu.vn. ORCID: https://orcid.org/0009-0002-9568-6342

Xuan-Truong Tran was born on December 22, 2004, in Thua Thien Hue. He is currently a fourth-year student in Information Technology, majoring in Computer Science at the University of Sciences, Hue University. Research areas: Artificial Intelligence, Computer Vision.
Email: 22T1020784@husc.edu.vn. ORCID: https://orcid.org/0009-0001-0804-5946

Dung Nguyen was born on June 13, 1988 in Thua Thien Hue. He graduated with a bachelor's degree in information technology from the College of Sciences, Hue University in 2010. In 2013, he graduated with a master's degree in computer science from the College of Sciences, Hue University. Currently he works at the University of Sciences, Hue University. Research fields: Software technology, artificial intelligence, machine learning, deep learning, databases.
Email: nguyendung@hueuni.edu.vn. ORCID: https://orcid.org/0009-0000-4510-7504. Tel: 0905198887.

Duc-Phuc Nguyen was born on January 2, 1987, in Hue City. He earned his bachelor's degree in Information Technology from the University of Sciences, Hue University in 2008, and his master's degree in Computer Science in 2017 from the same university. He has been working at the Department of Administration and Facilities of the University of Arts, Hue University since 2008. Research areas: Databases, Computer Networks.
Email: ndphuc.hufa@hueuni.edu.vn. ORCID: https://orcid.org/0009-0000-5916-9466