Regular Article

UIT-MLReceipts: A Multilingual Benchmark for Detecting and Recognizing Key Information in Receipts

Nguyen Tan Tran Minh Khang

University of Information Technology, Ho Chi Minh City, Vietnam

Correspondence: Nguyen Tan Tran Minh Khang, khangnttm@uit.edu.vn

Communication: received 21 December 2022, revised 21 December 2023, accepted 28 December 2023
Online publication: 03 January 2024, Digital Object Identifier: 10.21553/rev-jec.330
The associate editor coordinating the review of this article and recommending it for publication was Prof. Hoang Van Phuc.

Abstract– **The 4.0 industrial evolution has paved the way for development potential and revolution in Vietnam. In this movement, digitization appears to be necessary to transform numerous traditional economic sectors. It will provide valuable digital data for many automation applications and decision-making processes. Particularly in the retail industry, data has long played a vital factor. Hence, digitizing documents such as receipts can help businesses in management and enterprise development. Nevertheless, the digital transformation process is still slow because of the shortage of cleaned datasets for this type of document. This paper introduces a new dataset named UIT-MLReceipts for extracting key information in receipts. The task includes two sub-tasks: Receipt Text Detection (RTD) and Receipt Text Recognition (RTR). We thoroughly evaluate current state-of-the-art Receipt Text Detection using Faster R-CNN, YOLOv3, YOLOF, and Faster R-CNN with Precise RoI-Pooling on our dataset. To evaluate the performance of Receipt Text Recognition, we experiment with two text recognition baselines: RobustScanner and SATRN. Experimental results indicate that the Faster R-CNN with Precise RoI-Pooling outperforms the competitors and achieves the best mean Average Precision (mAP) score at 51.6% in the Receipt Text Detection task. With the Receipt Text Recognition task, results show that SATRN performs better. The dataset is available online for non-commercial research at http://tinyurl.com/uit-ml-receipts.**

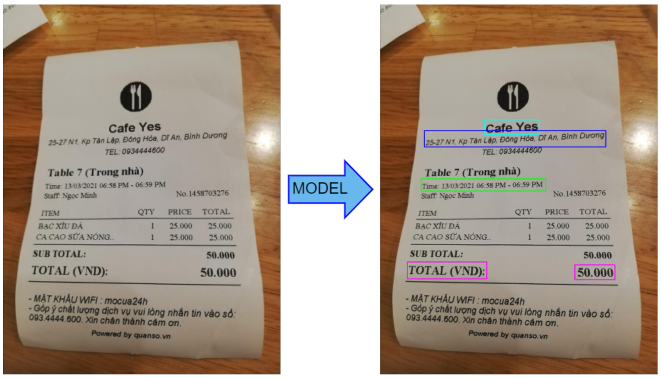
Keywords– **Receipt images, UIT-MLReceipts, object detection.**

1 Introduction

Digitization plays a crucial role in the 4.0 industrial revolution. Through digital transformation and information extraction, much data which used to be analog will be able to be stored and processed in computers. Hence, it will provide informative data for the social management of the government and push the development of many sectors. The information detecting problem takes an image as an input and returns the location and type of information in the image (Figure 1). The results of the problem can be used to analyze or fed as an input to other information understanding problems [1].

In many sectors, digitization and information detection are still open problems. Especially for receipt data, the biggest challenge is the diversity of the type and appearance of information. Besides, the variety of synonyms of the texts in the receipts also poses a noticeable challenge for extracting [2]. To solve the information diversity issues, in the scope of this research, we only focus on key information of the receipts. Following the MC-OCR challenge [3], we consider the key information as four types: seller's name, shop's address, buying time and total cost.

In Vietnam, due to the development of the business sectors, an increasingly enormous number of receipts have been produced in recent years. The amount of data in these receipts can be a precious resource for the business managers and government to analyze. However, there is a lack of Vietnamese receipts datasets used for the information detecting problem.



Cafe Yes
25-27 N1, Kp Tân Lập, Đông Hòa, Dĩ An, Bình Dương
TEL: 0934444600

Table 7 (Trong nhà)
Time: 13/03/2021 06:58 PM - 06:59 PM
Staff: Ngọc Minh No.1458703276

ITEM	QTY	PRICE	TOTAL
BẠC XỈU ĐÁ	1	25.000	25.000
CA CAO SỮA NÓNG..	1	25.000	25.000

SUB TOTAL: 50.000
TOTAL (VND): 50.000

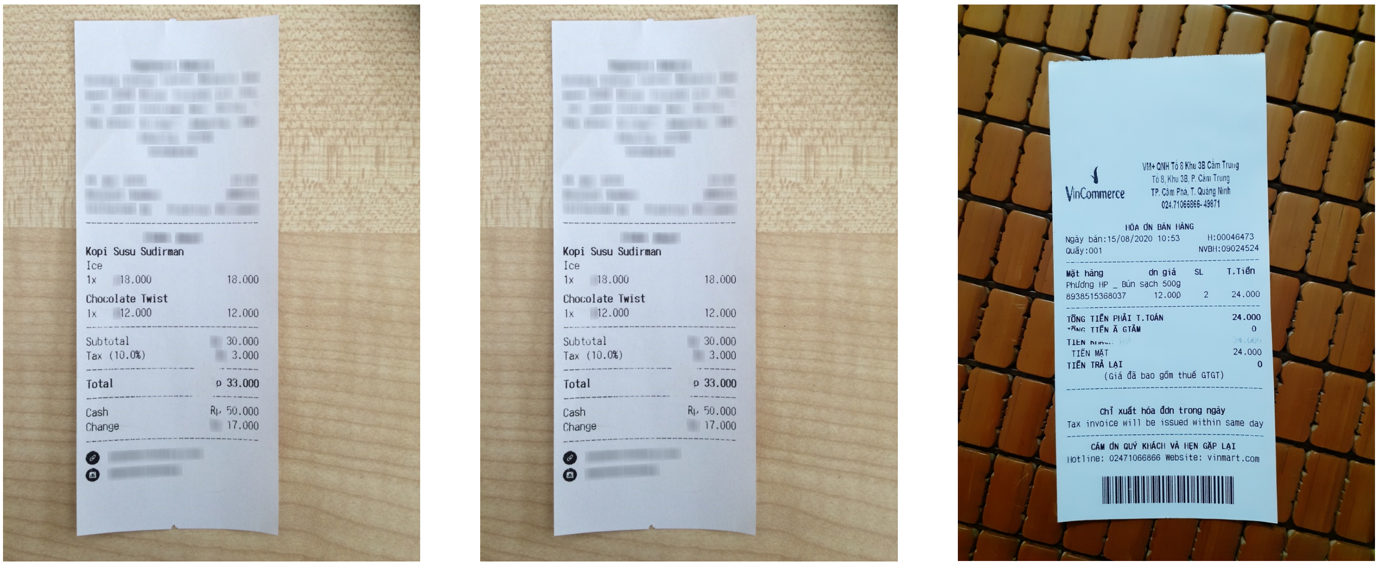
- MẬT KHẨU WIFI : mocua24h
- Góp ý chất lượng dịch vụ vui lòng nhắn tin vào số: 093.4444.600. Xin chân thành cảm ơn.

Powered by quanso.vn

Figure 1. The key information detection problem takes an image as an input and returns the location and type of key information in the image.

Thus, in this work, we introduce UIT-MLReceipts, a dataset for Vietnamese Receipt Text Localization and Receipt Text Recognition tasks. Our dataset consists of images of receipts from many kinds of retail stores, restaurants, and coffee shops in Vietnam. We hope that our dataset will contribute to larger studies in the area of Visual Document Understanding (VDU) [1]. Our main contributions to this paper include the following:

- We introduce a novel dataset for the task of detecting key information in the Vietnamese receipts.
- We have experimented on 3 well-known detectors including Faster R-CNN [6], YOLOv3 [7] and YOLOF [8] for detecting regions include key infor-


Kopi Susu Sudirman
Ice
1x 18.000 18.000
Chocolate Twist
1x 12.000 12.000
Subtotal 30.000
Tax (10.0%) 3.000
Total 33.000
Cash 50.000
Change 17.000
VinCommerce
VM+ QNH Tổ 8 Khu 3B Cẩm Trung
Tổ 8, Khu 3B, P. Cẩm Trung
TP. Cẩm Phả, T. Quảng Ninh
024.71066866-49871
HÓA ĐƠN BÁN HÀNG
Ngày bán:15/08/2020 10:53 H:00046473
Quầy:001 NVBH:09024524
Mặt hàng đơn giá SL T.Tiền
Phương HP _ Bún sạch 500g
8938515368037 12.000 2 24.000
TỔNG TIỀN PHẢI T.TOÁN 24.000
TỔNG TIỀN Ã GIẢM 0
TIỀN MẶT 24.000
TIỀN TRẢ LẠI 0
(Giá đã bao gồm thuế GTGT)
Chỉ xuất hóa đơn trong ngày
Tax invoice will be issued within same day
CẢM ƠN QUÝ KHÁCH VÀ HẸN GẶP LẠI
Hotline: 02471066866 Website: vinmart.com

(a) SROIE [4] (b) CORD [5] (c) MC-OCR challenge's dataset [3]

Figure 2. Sample images in 3 related datasets.

mations (Receipt Text Localization): seller, address, timestamp, and total cost.

- We present the Faster R-CNN using Precise RoI Pooling [9] instead of RoI Align, which achieves the best results among experiments.
- We experiment with the Receipt Text Recognition performance with two baselines: RobustScanner [10] and SAR [11].

The rest of the paper can be organized as follows. In section 2, we summarize the related works. Section 3 then describes the collecting, annotating process, and detailed information of our dataset. Section 4, we discuss the evaluation method and propose methods for our problem. Section 5, the evaluation and the outcomes obtained from different detection methods are presented. The paper ends with a conclusion and some directions for future work.

2 Related Work

2.1 Related Datasets

SROIE [4]. The SROIE dataset was published in the ICDAR2019 Robust Reading Challenge on Scanned Receipts OCR and Information Extraction. It includes 1,000 scanned images of English receipts in which there are about 600 train images and 400 test images, these images are all labeled (Figure 2a). Each image in the SROIE dataset will have two corresponding text file representing two types of annotations. The first type is used for the Scanned Receipt Text Localisation and Scanned Receipt OCR task. It contains bounding box information and a transcript of each line in the receipt. The other one is for Key Information Extraction from Scanned Receipts task which contains 4 information extracted from the receipt including company, date, address, total.

CORD [5]. CORD is also a dataset that was published in 2019. The CORD team found that although SROIE can be used to promote research in both OCR and information extraction tasks, it still have some limitations such as small data's volume, lack of bounding box for each word, etc. With the desire to build a larger, more effective and the first dataset with both box-level text and parsing class annotations, the authors released CORD. Its annotation has 3 main components: ***meta***, ***roi*** and ***valid_line***. ***meta*** contains general information of the image such as size, id, etc. ***roi*** contains information about the region of interest, specifically the bounding box coordinates of the receipt, and ***valid_line*** contains specific information about each line in the receipt along with the bounding box and transcript of each word in that line (Figure 2b). The dataset is expected to have up to 11,000 receipt images, but the author team has only published 1,000 sample images divided into 800 train images, 100 val images and 100 test images.

MC-OCR challenge's dataset [3]. The dataset of the RIVF2021 MC-OCR challenge includes 2,436 images of Vietnamese receipts. The dataset of the RIVF2021 MC-OCR contest includes Vietnamese invoices with more than 2,400 images, including more than 1500 fully labelled images. The labels of these images include information about bounding-box, segmentation, classification (SELLER, ADDRESS, TIMESTAMP and TOTAL_COST) and transcripts of the lines containing information about seller's name, shop's address, buying time and total cost (Figure 2c).

2.2 Object Detection Methods

In order to evaluate our dataset, we employ two-stage and one-stage advanced object detection methods: Faster R-CNN [6], YOLOv3 [7] and YOLOF [8]. In this section, we provide a brief overview of these object detectors.

2.2.1 Faster R-CNN: Reference [6] proposed by Ren et al. in 2016, Faster R-CNN is an enhanced version of Fast

Table I
The Statistics of Publicly Available Datasets

Dataset	Images	Number of categories	Coverage	Year
SROIE	1,000	4 classes	English scanned receipt images	2019
CORD	1,000	5 superclass and 42 subclass	English receipt images	2019
MC-OCR challenge's dataset	2,436	4 classes	English & Vietnamese receipt images	2021
UIT-MLReceipts (our)	**2,147**	**4 classes**	**English & Vietnamese receipt images**	**2022**

R-CNN [12] and it is a two-stage detector. This model is a single, unified network for object detection consisting of 2 modules: the Region Proposal Network module (RPN module [6]) and the Fast R-CNN detector. The Region Proposal Network will serve as the 'attention' of the system, telling the Fast R-CNN detector where to look. This module is also the main alteration of Faster R-CNN compared to its predecessor. The RPN module not only improves the system's speed but also helps Faster R-CNN achieve state-of-the-art detection accuracy on PASCAL VOC 2007 [13], 2012 [14], and MS COCO datasets. In addition, the Faster R-CNN model and RPN are the foundation of the first-place winning entries in ILSVRC [15] and COCO 2015 competitions.

2.2.2 YOLOv3: [7] despite the rise of its ancestor (YOLOv4 [16], YOLOv5), YOLOv3 is still a well-known one-stage detector, and it attained its popularity in many object detections related problems until recent years. Overall, the YOLO [17] model and its ancestor take an image as an input and return a 3D tensor. This tensor contains the partial information of the bounding boxes of the objects in the input image. Based on YOLOv2 [18], Redmon et al. [7] have made some modifications to build YOLOv3. First, YOLOv3 uses Darknet-53 [7], a larger version of Darknet-19 used in YOLOv2. In addition, the authors also add some shortcut connections which have not appeared in YOLOv2. Second, To help the model predict small objects better, the authors designed YOLOv3 to predict boxes in three different scales. These scales are obtained by downsampling the dimension of the input image by 32, 16, and 8, respectively. Finally, instead of using the softmax like its predecessors, YOLOv3 uses independent logistic classifiers in classification task. Moreover, the model also uses the binary cross-entropy loss for the training. The logistic classifiers help the model to solve the overlapping label challenge. Therefore, it will help YOLOv3 be suitable for many multilabel complex domains; in the regression task, the bounding box prediction methods of YOLOv3 follow the ones of YOLOv2. The model uses logistic regression to estimate the objectness score - the probability of an object surrounded by a bounding box.

2.2.3 YOLOF: [8] The YOLOF model is a simple and efficient baseline without the use of FPN (Feature pyramid networks [19]) presented in 2021. In this model, the authors try to use a simple Single-in-Single-out encoder instead of the complex Multiple-in-Multiple-out (MiMo) encoder (considered as the FPN). To bridge the gap between the Multi-in-Multi-out (MiMo) encoder and the Single-in-single-out (SiSo) encoder, the authors proposed two key components in YOLOF: Dilated Encoder and uniform matching. These components help YOLOF achieves comparable results with its feature pyramid counterpart RetinaNet [20] but 2.5× faster.

3 UIT-MLReceipts

3.1 Motivation

In the process of learning the datasets to train for the task of extracting receipts information, we found that besides the advantage of being fully and carefully labelled, existing datasets still have some drawbacks. Although SROIE [4] and CORD [5] are labelled very carefully, it only has 1,000 images. MC-OCR challenge's dataset is a large dataset of Vietnamese receipts. However, some images in the dataset are labelled with missing bounding boxes or wrong labels, wrong transcripts. Therefore, we publish the dataset UIT-MLReceipts with the desire to enrich the existing data source.

3.2 Data Preparation

To get the diversity of the structure, colour, font, format of the receipts, we collect them from restaurants, cafes, bookstores, convenience stores, supermarkets at many different places. Collected receipts are captured by our mobile phone on many backgrounds with different angles and brightness to raise variety to the dataset 3.

Besides, we also collect some receipt images from travel and experience sharing groups on social networks. Receipt images from social networks also have a variety of angles and sizes.

After collecting receipt images, we proceed to remove receipt images that contain less than two types of information objects to be extracted or receipt images that are too blurry and can't see the text clearly to increase the efficiency of the model training process. After preparing data, we obtained 2,147 receipt images for the new dataset. We divided these images into three parts, including 1,000 training images, 358 validation images and 789 test images.

3.3 Annotation Pipeline

In the data labelling step, we used the Faster R-CNN model that has been trained through the RIVF2021 MC-OCR challenge's dataset to predict the bounding box of the key information (store name, store address, printing time, total amount). We then manually edited these bounding-boxes with some rules as shown below:

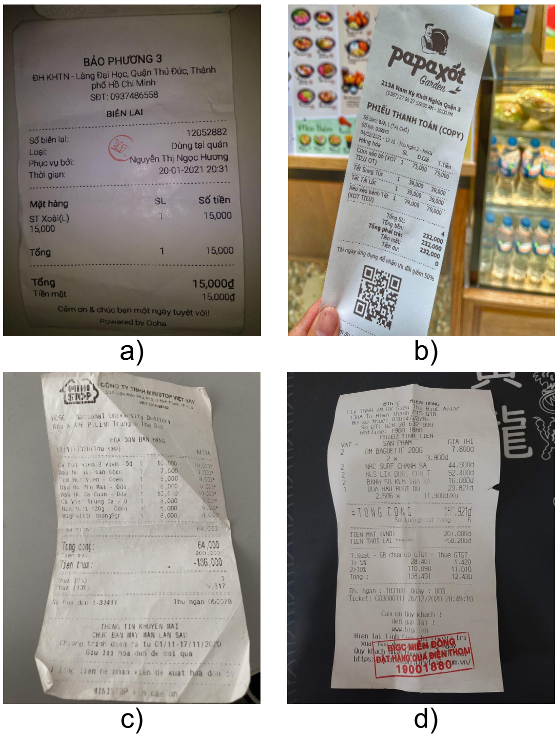

a) b) c) d)

Figure 3. Some images showing the variety of receipt structure and how it is photographed in the dataset.

- Each object consists of only one line, the dropped information will be assigned to another object.
- If the words are too far apart, we will split them into two different ones.
- **SELLER:** name of store, or branch name without house number, company name if it is bold or contains store name.
- **ADDRESS:** address (building name, house number, street name, etc.) of the store, or branch name that contains the structure of *house number + street name*.
- **TIMESTAMP:** receipt printed time, or the time the customer leaves if there is no receipt printed time and corresponding keywords.
- **TOTAL_COST:** the total amount the customer has to pay for the receipt (after discount) and the corresponding keyword.

After this process, we got the complete and accurate bounding box of the information that needs to be extracted from all 2,147 images of the dataset with COCO format [21].

3.4 Dataset Description

UIT-MLReceipts is one of the largest datasets of receipts in Vietnam, with 2,147 receipt images. We annotated the images with bounding boxes and classified them into four classes of objects commonly found in the invoices: SELLER, ADDRESS, TIMESTAMP, TOTAL_COST. The statistic of these labels is showed in Figure 4. Receipts in the dataset were collected from restaurants, cafes, bookstores, grocery stores, and supermarkets in Vietnam. Besides, we added some images from groups sharing about travels, experiences on social networks. The dataset is available for access through the provided link[1]. This dataset has the following special features:

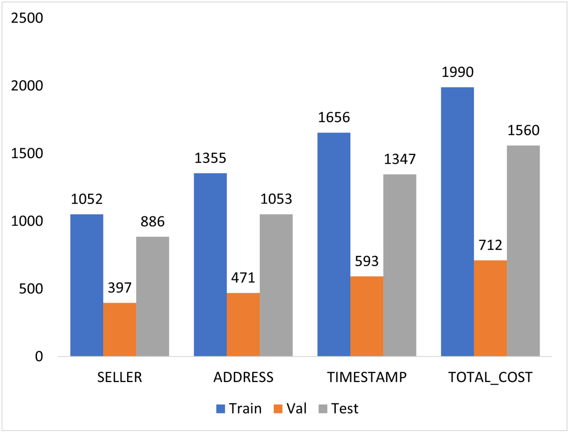


Figure 4. Statistical of our dataset.

Diversity of receipt characteristics. The receipts in this dataset are collected from many different stores, so they have differences in colour, information structure, font, font size, spacing between sections, information display.

Diversity of languages. Most receipts in Vietnam use Vietnamese with or without accents, some use English receipts, a few use a combination of Vietnamese and English (Figure 5 show statistics of the main languages used in the receipts of the dataset). This promises to be a challenge.

Diversity of status of receipts. Due to being printed from different printers, some receipts is more blurred than the others; others are printed incorrectly, causing the loss of characters. Moreover, we noticed that some samples were dirty, bent, wrinkled, torn, holed, stamped. These issues lead to unclear words and the deformation of some lines.

Diversity in image characteristics. The receipt photos in the dataset are taken from different phones, with varying angles of shooting and lighting conditions, so they have different sizes, resolutions, brightness, receipt shot angle.

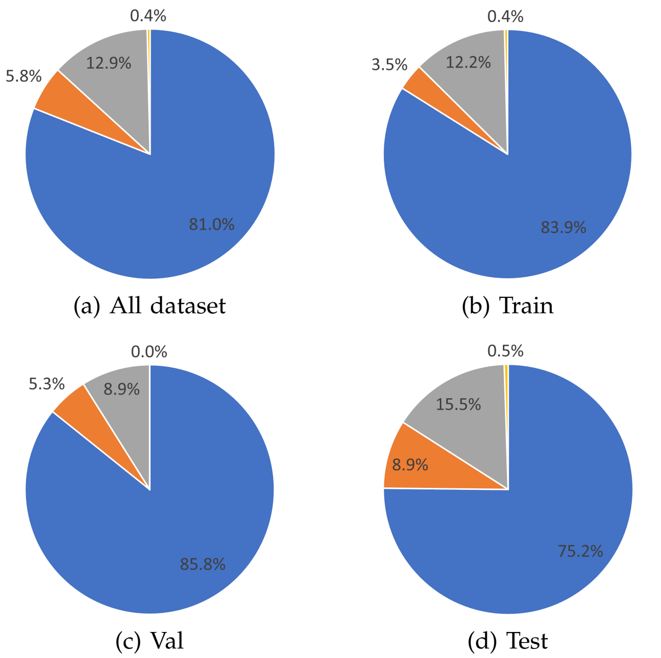
4 Precise Faster R-CNN

In an attempt to improve Faster R-CNN's performance for regions of interest detection on our dataset UIT-MLReceipts, we present the Precise Faster R-CNN method, which replaces the default RoI Align with Precise RoI Pooling for extracting the objects' features to fix-sized vectors. This section clearly describes the Precise RoI Pooling and how we apply it in Faster R-CNN.

4.1 Precise RoI Pooling

Precise RoI Pooling (PrRoI) was proposed by Jiang, Luo, Mao, Xiao, and Jiang [9] to avoid quantization

[1] http://tinyurl.com/uit-ml-receipts



(a) All dataset

(b) Train

(c) Val

(d) Test

Figure 5. Statistics of the main languages used in the receipt of the dataset include: Vietnamese with accents (blue), English (orange), Vietnamese without accents (gray), other languages (yellow).

which appeared in normal RoI Pooling and RoI Align. First, given the feature map F before applying RoI Pooling, $w_{i,j}$ is the value at (i, j) position on the feature map, the discrete feature map could be considered continuous at any continuous coordinates via bilinear interpolation function below

$$f(x, y) = \sum_{i,j} IC(x, y, i, j) \times w_{i,j}, \tag{1}$$

where $IC(x, y, i, j) = \max(0, 1 - |x - i|) \times \max(0, 1 - |y - j|)$ is defined as interpolation coefficient. Then, the coordinates of top-left and bottom-right points of RoI (Region of Interest) are denoted (x_1, y_1) and (x_2, y_2), respectively, which become the continuous coordinates in Equation 1. The RoI feature is now extracted from the original feature map via the PrRoI Pooling module, which can be expressed by the Equation 2 below

$$\text{PreciseRoIPooling}(\text{RoI}, \text{F}) = \frac{\int_{y_1}^{y_2} \int_{y_2}^{y_1} f(x, y) dxdy}{(x_2 - x_1) \times (y_2 - y_1)}. \tag{2}$$

4.2 Applying Precise RoI Pooling to Faster R-CNN

Faster R-CNN commonly uses RoI Align and RoI Pooling as RoI poolers to obtain fix-sized objects' vectors in default architecture. These methods cause quantization, which quickly loses the critical information of salient objects, such as borders, aligns, etc. Therefore, besides running the experiment with Faster R-CNN, we also do an experiment that replaces the default RoI Align with Precise RoI Pooling to obtain the quality-higher fix-sized vectors for further regression and classification tasks.

5 Text Recognition Methods

For evaluating the Receipt Text Recognition task, we use two baselines: RobustScanner [10], and SAR [11] methods. These two baselines are LSTM-based models, which take inputs as visual features extracted from CNN backbones, and auto-aggressively recognize the texts using LSTM layers.

5.1 RobustScanner

Yue, Kuang, Lin, Sun, and Zhang propose RobustScanner based on adding a new branch combined with the decoder's attention module. RobustScanner consists of one encoder and one decoder. In the encoder, RobustScanner adapts a 31-layer ResNet as a backbone. The decoder is made up of one hybrid branch, one position enhancement branch, one dynamically-fusing module, and one prediction module. The hybrid branch consists of a two-layer LSTM of 128 hidden state sizes and an attention module. The LSTM takes the previously predicted character as input and generates the query vector h_t. Then the query vector is fed into the attention module to estimate the glimpse vector g_t for character prediction during decoding. The position enhancement branch is designed to mitigate the problem that context cannot be reliably used to predict characters because the positional information becomes weak. In contrast, the contextual information becomes strong during decoding. This proposed branch consists of a position embedding layer, a position-aware module, and a attention module. The position embedding layer encodes the decoding time step. The Position Awareness Module is proposed to capture global and high-level information in order to the encoder output feature map is position-aware. Two-layer LSTM has 128 hidden state sizes for each row of the feature map F to capture the global context. Finally, dynamically fuse the hybrid brand outputs g_t and the output of the position enhancement branch g'_t at each time step t. Additionally, RobustScanner contains a gate mechanism to predict an attention weight for each dimension of their concatenation which is used to enhance or suppress their similarity feature.

5.2 SATRN

SATRN is the model based on Transformer architecture, including Transformer Encoder and Transformer Decoder. The encoder is used to learn the high-level relation representation between tokens in the feature maps, and the decoder decodes them into characters. Li, Wang, Shen, and Zhang observed that previous studies fit 1D feature vectors to learn text recognition is not practical. Because the regions of text that appear in the image have various shapes (heavily curved or rotated), 1D feature vectors can not reflect shape information in the model space even if the positional encoding is used [22–24]. Therefore, SATRN introduced the effective 2D positional encoding A2DPE, which is also the key component. The proposed scheme can be formulated as follow equations

$$\mathbf{p}_{hw} = \alpha(\mathbf{E})\mathbf{p}_{h}^{sinu} + \beta(\mathbf{E})\mathbf{p}_{w}^{sinu}, \quad (3)$$

where $\mathbf{p}_{h}^{sinu}$ and $\mathbf{p}_{w}^{sinu}$ are sinusoidal positional encoding over height and width, respectively. Operations to calculate $\mathbf{p}_{h}^{sinu}$ and $\mathbf{p}_{w}^{sinu}$ were introduced by Vaswani et al. [25]. $\mathbf{E}$ denotes the 2D input features obtained from a shallow convolutional neural network backbone. $\alpha(\cdot)$ and $\beta(\cdot)$ are simple networks including two fully connected layers to compute scale factors for height and width, which can be formulated as following equations:

$$\alpha(\mathbf{E}) = Sigmoid(Max(0, Pooling(\mathbf{E}\mathbf{W}_{1}^{h}))\mathbf{W}_{2}^{h}), \quad (4)$$

$$\beta(\mathbf{E}) = Sigmoid(Max(0, Pooling(\mathbf{E}\mathbf{W}_{1}^{w}))\mathbf{W}_{2}^{w}), \quad (5)$$

where $\mathbf{W}_{1:2}^{h,w}$ denotes the linear projections. *Pooling* denotes the average pooling. By learning α and β, the Transformer model can adaptively adjust the length along with height and width.

6 Experiment

6.1 Experimental Setting

In this paper, we conduct experiments on GeForce RTX 2080 Ti GPU with 11019 MiB. We implement Faster R-CNN, YOLOv3 and YOLOF on the MMDetection [26] toolbox V2.18.1. Faster R-CNN detector is trained with default configuration within 12 epochs, and ResNet-101 architecture is used as the backbone for feature extraction. With the YOLOF model, we use the ResNet50 backbone for feature extraction and also train within 12 epochs. The YOLOv3 detector requires more training time, costing 273 epochs; the training configuration is kept as default. To train the RobustScanner and SATRN models, we use MMOCR toolbox [27]. We keep the default configuration as reported in the original paper. With RobustScanner, we use the ResNet31 backbone to extract the feature maps from images. With SATRN, the same shallow CNN model in the study [11] is used. The maximum width and length of images are set as 100.

6.2 Evaluation Metrics

Receipt Text Localization (RTL). To evaluate the performance of object detection methods, we calculate the Average Precision metric using COCO API [2]. We calculate the AP scores of all classes and take the average of them as the mean AP score (mAP). This process could be briefly expressed by the Equation 6 and 7:

$$\mathrm{AP_c} = \frac{1}{\#T} \sum_{IoU \in T} AP[c, IoU], \quad (6)$$

$$\mathrm{mAP} = \frac{1}{\#C} \sum_{c \in C} \mathrm{AP_c}, \quad (7)$$

where AP_c is the Average Precision of c-th class; C is the set of all classes in the dataset; T is the set of IoU threshold $T \in [0.5 : 0.05 : 0.95]$. Besides, we also calculate the mAP scores at IoU = 0.5 and IoU = 0.75, which called AP@50 and AP@75, respectively.

Receipt Text Recognition (RTR). To evaluate the performance of text recognizers on our UIT-MLReceipts dataset, we use three metrics: Recall, Precision (on character level), 1-N.E.D, and CER. Recall and Precision in this problem can be formulated as Equation follows:

$$TP = NumMatch(s^{p}, s^{gt}), \quad (8)$$

$$Recall = \frac{TP}{Length(s_{gt})}, \quad (9)$$

$$Precision = \frac{TP}{Length(s_{p}^{*})}, \quad (10)$$

where TP is the number of true positive predictions. $Length(\cdot)$ is the operation that calculates the length of ground truth or predicted texts. $NumMatch(\cdot)$ is the operation that calculates the number of the same characters between two sequences. s^{p}, s^{gt} is the predicted text and ground-truth text, respectively.

Besides, 1-N.E.D can be formulated as following Equation

$$Norm = 1 - \frac{1}{N} \sum_{i=1}^{N} D(s_{i}^{gt}, s_{i}^{p}) / Max(s_{i}^{gt}, s_{i}^{p}), \quad (11)$$

where $D(\cdot)$ is the Levenshtein Distance [cite], N is the number of samples in the test set of our UIT-MLReceipts dataset.

Finally, the CER score is calculated as following Equation

$$CER = \frac{1}{L} \sum_{1}^{N} (i + s + d), \quad (12)$$

where $\sum_{1}^{N}(l_i)$ is the total length of all reference texts of the test set; l_i is the length of i^{th} document; N is the total number of text samples; $(i + s + d)$ corresponds to the minimal numbers of character insertions i, substitutions s and deletions d required to transform the reference text into the OCR output.

6.3 Experimental Results and Discussion

Performance of Receipt Text Detection. Following the aforementioned experiment setting, we intensively experiment with four state-of-the-art methods: Faster R-CNN [6], YOLOv3 [7], YOLOF [8] and Precise Faster R-CNN. TABLE II shows the performance comparison of these methods on UIT-MLReceipts. Overall, our PrRoI Pooling replacement positively affects the Faster R-CNN model with competitive results, compared to the current state-of-the-art methods on UIT-MLReceipts. Thanks to PrRoI, bounding boxes with large excess outside the object will be removed (TOTAL_COST object in Figure 6) and thereby improving model efficiency. Out of the other methods, the result of Precise Faster

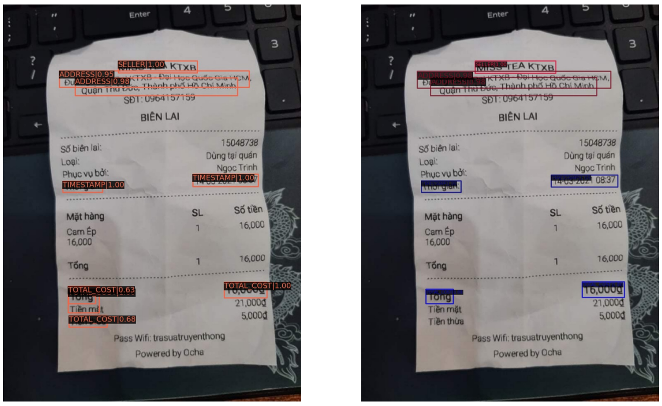
[2] https://github.com/cocodataset/cocoapi

Table II
Experimental Results on Regions of Interest Detection (%)

#	Methods	SELLER	ADDRESS	TIMESTAMP	TOTAL_COST	AP	AP@50	AP@75
1	YOLOF	27.7	9.1	26.8	33.3	24.2	5.71	14.9
2	YOLOv3	48.8	37.9	45.8	**52.5**	46.3	82.4	49.1
3	Faster R-CNN	53.5	**53.9**	45.0	48.7	50.3	81.2	59.1
4	Precise Faster R-CNN	**54.8**	52.7	**48.1**	50.9	**51.6**	**82.6**	**61.6**

Table III
Experimental Results of Receipt Texts Recognition Task Using RobustScanner and SATRN on UIT-MLReceipts Dataset

Methods	Type of Information	Recall↑	Precision↑	1-N.E.D↑	CER↓
RobustScanner	SELLER	0.5696	0.5599	0.4742	0.5946
	ADDRESS	0.3766	0.4515	0.3283	0.7243
	TIMESTAMP	0.5934	0.5863	0.5629	0.5126
	TOTAL COST	0.7485	0.6994	0.6911	0.3930
	Average	0.5720	0.5743	0.5141	0.5561
SATRN	SELLER	0.6074	0.6507	0.5483	0.4845
	ADDRESS	0.4226	0.5918	0.4082	0.6308
	TIMESTAMP	0.6586	0.6806	0.6414	0.3903
	TOTAL COST	0.8707	0.8665	0.8535	0.1740
	Average	**0.6398**	**0.6974**	**0.6129**	**0.4199**



(a) Roi-Pooling (b) Precise Roi-Pooling

Figure 6. Comparison experimental results between RoI-Pooling and Precise RoI-Pooling.

R-CNN is highest at 54.8% mAP. The Precise Faster R-CNN achieves the highest performance in SELLER, and TIMESTAMP objects at 54.7% and 48.1%, respectively. However, for ADDRESS object detection, the baseline Faster R-CNN obtains the higher score at 53.9%. On the other hand, YOLOF and YOLOv3 have significantly lower scores than the Faster R-CNN and Precise Faster R-CNN with 24.2% and 46.3% mAP average scores and on every considered object except TOTAL_COST object. These results imply that in the receipt's key information detecting problems, the two-stages methods are more robust than the one-stage ones.

Performane of Receipt Text Recognition. For solving the Receipt Text Recognition task, we train RobustScanner and SATRN model. The experimental results are reported in Table III. It can be seen that SATRN performs significantly better than RobustScanner. SATRN achieves 0.6398, 0.6974, 0.6129, and 0.4199 on Recall, Precision, 1-N.E.D, and CER scores, respectively, on average for four types of information. However, UIT-MLReceipts is quite challenging for text recognition models for SELLER and ADDRESS; the text recognizers do not perform well on these types of information. In detail, RobustScanner returns 0.5946 CER on SELLER and 0.7243 on ADDRESS, which is exceptionally high. SATRN shows better results, 0.4845 CER on SELLER and 0.6308 on ADDRESS, but these numbers are still quite modest. The reason is that various font styles on Vietnamese receipts are captured in challenging conditions: blurred and slanted. Moreover, Vietnamese has more characters than English because it uses UTF-8 characters beside Latin to present accents. Additionally, SELLER and ADDRESS are two types of information that include more diverse characters than others. It forces models to learn to recognize more characters and causes difficulty in inference time. By observing the experimental results, our UIT-MLReceipts dataset raises the high demand for developing better text recognition models to adapt well to the Receipt Text Recognition task.

7 Conclusion

In this paper, we publish UIT-MLReceipts - a new dataset for detecting key information in receipts. The dataset has 2,147 receipt images which are taken in many different backgrounds and have many different structures. With the proposed dataset, we evaluate and analyze state-of-the-art methods on two tasks: Receipts Text Localization (RTL) and Receipts Text Recognition (RTR). Experimental result shows that our proposed Precise Faster R-CNN achieves the best performance at 54.8% mAP score in the RTL task, and SATRN is the baseline model that obtains the highest scores in all metrics for the RTR task.

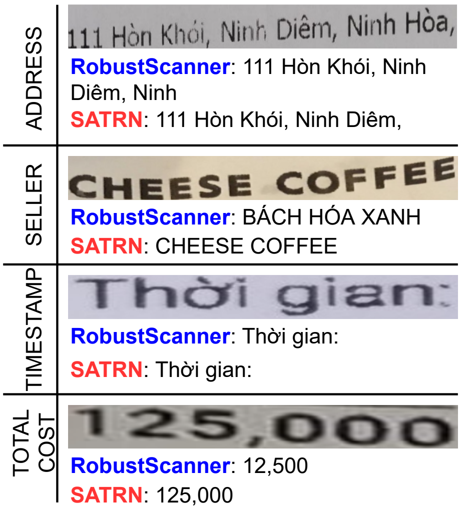

ADDRESS
111 Hòn Khói, Ninh Diêm, Ninh Hòa,
RobustScanner: 111 Hòn Khói, Ninh Diêm, Ninh
SATRN: 111 Hòn Khói, Ninh Diêm,
SELLER
CHEESE COFFEE
RobustScanner: BÁCH HÓA XANH
SATRN: CHEESE COFFEE
TIMESTAMP
Thời gian:
RobustScanner: Thời gian:
SATRN: Thời gian:
TOTAL COST
125,000
RobustScanner: 12,500
SATRN: 125,000

Figure 7. Comparison experimental results of Receipt Text Recognition between RobustScanner and SATRN

References

[1] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "Docformer: End-to-end transformer for document understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 993–1003.

[2] D. C. Bui, D. Truong, N. D. Vo, and K. Nguyen, "MC-OCR Challenge 2021: Deep Learning Approach for Vietnamese Receipts OCR," in *Proceedings of the 2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 2021, pp. 1–6.

[3] X.-S. Vu, Q.-A. Bui, N.-V. Nguyen, T. T. H. Nguyen, and T. Vu, "MC-OCR Challenge: Mobile-Captured Image Document Recognition for Vietnamese Receipts," in *Proceedings of the 2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 2021, pp. 1–6.

[4] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. Jawahar, "ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction," in *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1516–1520.

[5] S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, and H. Lee, "CORD: a consolidated receipt dataset for post-OCR parsing," in *Proceedings of the Workshop on Document Intelligence at NeurIPS 2019*, 2019.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Proceedings of the Advances in neural information processing systems*, 2015, pp. 91–99.

[7] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *ArXiv*, vol. 1804.02767, 2018.

[8] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 039–13 048.

[9] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 784–799.

[10] X. Yue, Z. Kuang, C. Lin, H. Sun, and W. Zhang, "Robustscanner: Dynamically enhancing positional clues for robust text recognition," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 135–151.

[11] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8610–8617.

[12] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://host.robots.ox.ac.uk/pascal/VOC/voc2007/.

[14] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://host.robots.ox.ac.uk/pascal/VOC/voc2012/.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[16] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *ArXiv*, vol. 2004.10934, 2020.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[18] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European conference on computer vision*. Springer, 2014, pp. 740–755.

[22] J. Wang and X. Hu, "Gated recurrent convolution neural network for ocr," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-Cross Attention for Semantic Segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[24] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "MMDetection: Open MMLab Detection Toolbox and Benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[27] Z. Kuang, H. Sun, Z. Li, X. Yue, T. H. Lin, J. Chen, H. Wei, Y. Zhu, T. Gao, W. Zhang *et al.*, "MMOCR: a comprehensive toolbox for text detection, recognition and understanding," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3791–3794.

Nguyen Tan Tran Minh Khang received his B.S degree and M.S degrees in Computer Science from University of Science, VNUHCM, Vietnam in 1990 and 1995. He received his Ph.D. degree in 2012 from the University of Science, VNUHCM, Vietnam. Currently, he is the Vice-President of University of Information Technology, VNUHCM, Vietnam. His research interests include artificial intelligence and computer vision.