FVI: An End-to-end Vietnamese Identification Card Detection and Recognition in Images

Hoang Danh Liem*, Nguyen Duc Minh*, Nguyen Bao Trung*, Hoang Tien Duc*,
Pham Hoang Hiep*, Doan Viet Dung* and Dang Hoang Vu*
* FPT Technology Research Institute
FPT Corporation

***Abstract*—The neccesity of digitializing all old constrained forms such identification card or register book have become a critical issue due to the importance of practical applications for sales or financial services. To address this issue, we develop an End-to-end Identification Card Recognition system which allows us to quickly detect, recognize text and extract important information from the ID card. We not only present the modeling technique for efficient detection and recognition of texts but also the architecture design of FVI which is currently deployed in several organizations. We performed extensive evaluations of the designed system as the verification of our efficient system for a large-scale detecion and recognition of constrained forms.**

***Index Terms*—Optical Character Recognition, text detection, text recognition**

I. Introduction

One of the key importance of machine learning in computer vision is to assist people to save time to extract important information from paper-only documents. Specially, with Vietnamese ID cards or register books which are normally in constrained forms and various conditions. One of the challenges in the image understanding problem called Optical Character Recognition (OCR). This process consists of conversion of digital image containing printed or typeset texts into machine-encoded text. The difficulties of recognizing typeset or printed texts with high level of noise in the background and intensity of text color is approximately identical to the background. Furthermore, another aspect of the problem is the huge amount of human-labeled data needed to train these models and the processing time required in the practice is also considered carefully.

In this paper, we have presented, implemented and deployed our scalable OCR system which primarily focus on detecting and recognizing Vietnamese ID cards. Our contributions consist of:

- The first end-to-end Vietnamese ID Detection and Recognition System which can recognize both old and new Vietnamese identification cards from any raw images shot by any decent smartphone cameras which contain one ID card. Figure 2 illustrates some results produced by our FVI system.
- We optimize the OCR module which utilizes curriculum learning schedule [1] for recognizing Vietnamese words.
- Our system is easily applicable to others constrained documents such as register books, driving license, etc.

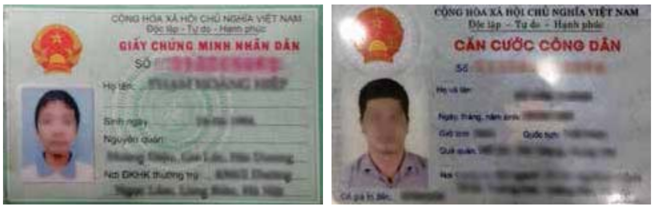

Fig. 1. Example of 2 types of a Vietnamese ID card

The rest of the paper is organized as follow: *II* provides overview and previous research of the OCR. *III* provides the architecture detail and dataflow of FVI. *IV and V* describe the component detail and the experimental evaluation of FVI.

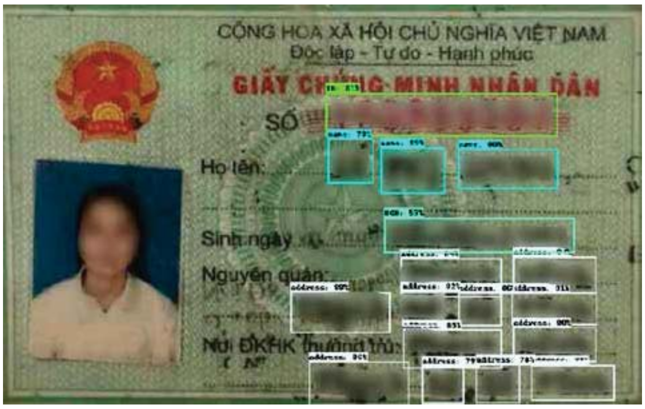

Fig. 2. OCR text recognition of the FVI. An approach based on RetinaNet [2] detects the individual words, and a fully-convolutional network produces transcription of each word.

II. Related Work

There are several OCR systems which detect and recognize natural scene-texts (text-in-the-wild) or on specific images.

For a quite remarkable text recognition model, CHAR model [3] tried to solve the recognition of natural scene-text problem which did not require any human-labeled data to train the model. This problem is reduced to three different classification problems, 90-thousand-word classification, character-based classification and bag-of-N-gram classification. Despite the fact that their models could perform the text recognition task at the whole-word level hollistically, all of the three approaches did not take advantage of sequential characteristic of a word as a sequence of characters even if Vietnamese words in this case. Furthermore, with restriction of word length, their models could not read any words which are more than 23 characters long.

In terms of text detection, one of the most widely-used models for both natural scene-text and document form-text is CTPN [4]. The CTPN is basically a combination of a typical deep CNN (e.g. VGG, ResNet) as a feature extractor and a BiLSTM as a sequential text connector. This model supposedly bridges the gap between general object detection and text detection and usually detects whole lines of text as it consists of as many as nearby words. In spite of fine-scale text proposals output by CTPN, the whole detected line from Vietnamse ID card is inappropriate for the immediate text recognition afterwards which may require an extra model in the pipeline to separate each word from the others.

The latest research that is up-to-date as the end-to-end detection and recognition of texts in images is Rossetta [5]. The large-scale system for OCR. They utilized Faster-RCNN [6] as the detection model and recurrent CTC loss [7] as the recognition method. Their system is efficient for just only little text on image but is not applicable to constrained forms with a lot of noise. Furthermore, their recognition module works well solely with English words and has a few limitations with Vietnamese words. We have adapted one of their concepts regarding curriculum learning schedule as the backbone of our Visual Attention-based recognition text.

III. System Architecture:

We describe our FVI system, which is deployed in several organizations. The system processes images which are uploaded to the system to assist salesperson to quickly obtain the important information such as ID number, full name, date of birth and address from shoppers. The image processing process within FVI contains the following steps with the dataflow in figure 4:

- Image is uploaded by client to the FVI system to be preprocessed and prepared for the batch-processing queue (Step 1 and 2).
- The ID localization module detects, crops the ID cards, and rotates it with the correct orientation (Step 3).
- The text detection module localizes bounding boxes for each interested words just within the cropped ID card ifself (Step 4).
- Each word is cropped along with its localized bounding box to form an individual image which contains only one according word, then the word image is passed to the text recognition module to extract information to text (Step 5).
- The information is transferred back to the front-end server and returned results to the client (Step 6 and 7).

IV. Component Detail

In this section, we describe each main module of our FVI system in detail.

A. ID Card Cropping

In order to identify the ID card correctly, we need to localize where the ID card is. To obtain that, we applied a simple object detection algorithm to identify 4 corners of the ID card based on the model [2]. Although, there are many traditional computer vision methods can be utilized to crop the ID card as a rectangle for example Hough Transform, etc. However, they are not applicable to many variations of the Vietnamese ID card which are usually degraded with high level of noise.

B. Text Extraction Model

We perform OCR in 2 independent steps: detection and recognition. In the first step, we detect rectangular regions potentially containing text in the image. In the next step, we perform text recognition task where Inception-v3 network is used to extract features and Visual Attention mechanism [8] is used to read the word.

Text Detection: We adapted an approach based on RetinaNet [2], as a state-of-the-art one stage detector. Overall, RetinaNet performs detection task by: i) Learning a Feature Pyramid Network (FPN) [9] combined with the backbone feature extractor ResNet50 [10] that can present an image as a multi-scale convolutional feature maps. ii) Learn a region based on translation-invariant anchor boxes which takes the input from multi-scale feature maps and produce the region boxes that contain text with probabilities. iii) 2 subnets as a box regressor and a object classifier to detect them.

Text Recognition Model: We followed the approach from Google reseachers [8]. First, we utilized the Inception-v3 CNN network for feature extraction. The output of the feature extraction phase is passed to the RNN network. Our RNN network is basically a BiLSTM which is combined with the Spatial Attention to predict the mask based on the current RNN state. Different from the original paper, we adapted the cirriculum learning technique as training schedule from [5]. At the first 10 epochs, we warmed up the training with very short words. We started training with words of lengh $<= 3$, which is considered easy and quickly alignment and where the variations in length is tiny. We increase the maximum length of word at every epoch. Simutaneously, we started with a very small learning rate and keep increasing the learning until it reaches our initial learning at 1.0.

V. Experiment

A. Performance on text recognition

Based on the requirement of FVI structure, we have several metrics to evaluate the performance on the whole OCR pipeline: ID number, name, date of birth and address. We performed our experiment on 1000 images which contain ID cards with arbitrary positions to obtain the text. Afterwards, we compare our results with groundtruths to decide the detection and recognition is correct or incorrect. The result of the system is shown in the table **I**.

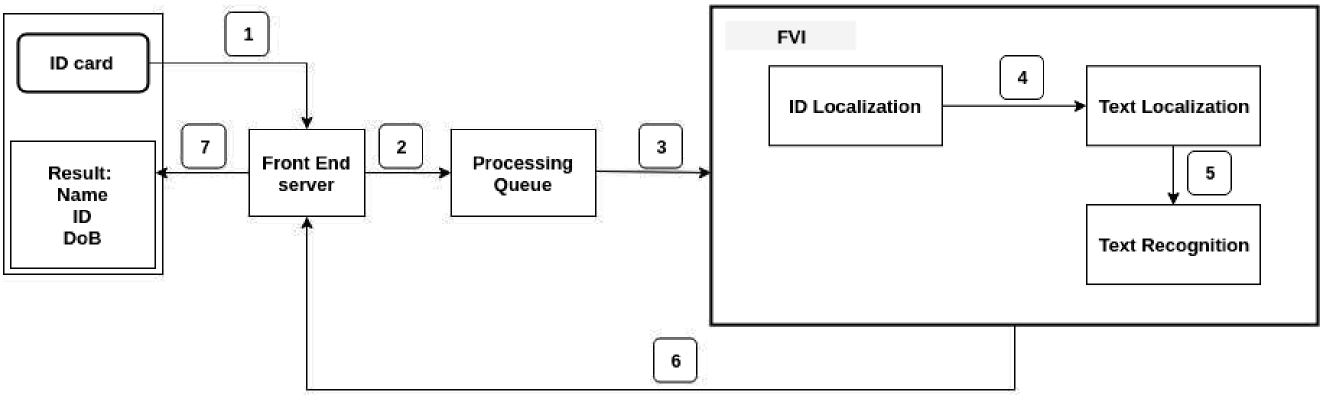

Fig. 3. Architecture of FVI

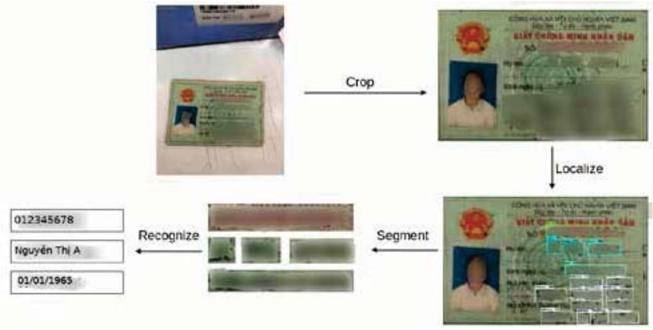

Fig. 4. Dataflow of FVI

Metric	Accuracy (%)
ID number	86.7 %
Name	85.7%
DoB	79.7%
Address	70.24%
ID number + Name+ DoB	68.5%

TABLE I
ACCURACY OF TEXT RECOGNITION

Platform	Inference Time (s)
CPU	12.4234
GPU	1.6456

TABLE II
INFERENCE TIME BASED ON THE PLATFORM

B. Inference time:

We perform the inference on both GPU and CPU on a workstation. The workstation specification consists of CPU Intel Core i5 7400, memory 16GB DDR4 and GPU NVIDIA GTX 1060. We run the inference of 100 different images which is in the batch size of 100. Afterwards, we calculate the average time over 100 images. The result is shown on the Table II.

ACKNOWLEDGMENT

We would like to thank all co-workers from FTRI: FPT Technology Research Institute and others who contributed, supported and collaborated with us during the development, experiment, and deployment of our system.

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, 2009. ACM.

[2] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollr. Focal loss for dense object detection. PP:1–1, 07 2018.

[3] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. 06 2014.

[4] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9912 LNCS:56–72, 2016.

[5] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large Scale System for Text Detection and Recognition in Images. *Kdd*, pages 71–79, 2018.

[6] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.

[7] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 369–376, New York, NY, USA, 2006. ACM.

[8] Zbigniew Wojna, Alexander N. Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. Attention-based extraction of structured information from street view imagery. pages 844–850, 11 2017.

[9] Tsung-Yi Lin, Piotr Dollr, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. 12 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.