Computer Systems Science & Engineering
DOI:10.32604/csse.2022.019064
Article

Tech Science Press

An improved CRNN for Vietnamese Identity Card Information Recognition

Trinh Tan Dat[1], Le Tran Anh Dang[1,2], Nguyen Nhat Truong[1,2], Pham Cung Le Thien Vu[1], Vu Ngoc Thanh Sang[1], Pham Thi Vuong[1] and Pham The Bao[1,*]

[1]Information Science Faculty, Sai Gon University, HCM City, 700000, Vietnam
[2]Faculty of Electrical & Electronics Engineering, University of Technology, HCM City, 700000, Vietnam
[*]Corresponding Author: Pham The Bao. Email: ptbao@sgu.edu.vn
Received: 31 March 2021; Accepted: 10 May 2021

Abstract: This paper proposes an enhancement of an automatic text recognition system for extracting information from the front side of the Vietnamese citizen identity (CID) card. First, we apply Mask-RCNN to segment and align the CID card from the background. Next, we present two approaches to detect the CID card's text lines using traditional image processing techniques compared to the EAST detector. Finally, we introduce a new end-to-end Convolutional Recurrent Neural Network (CRNN) model based on a combination of Connectionist Temporal Classification (CTC) and attention mechanism for Vietnamese text recognition by jointly train the CTC and attention objective functions together. The length of the CTC's output label sequence is applied to the attention-based decoder prediction to make the final label sequence. This process helps to decrease irregular alignments and speed up the label sequence estimation during training and inference, instead of only relying on a data-driven attention-based encoder-decoder to estimate the label sequence in long sentences. We may directly learn the proposed model from a sequence of words without detailed annotations. We evaluate the proposed system using a real collected Vietnamese CID card dataset and find that our method provides a 4.28% in WER and outperforms the common techniques.

Keywords: Vietnamese text recognition; OCR; CRNN; BLSTM; attention mechanism; joint CTC-Attention

1 Introduction

In computer vision, scene text recognition (STR) related to analyzing and understanding high-level semantic information from texts in the image is challenging. The STR systems have been extensively and successfully utilized in various applications, such as image retrieval, driver-assisted systems, recognition of personal cards and related documents, etc. [1–2]. The STR systems often include two main procedures: Text detection and recognition. Text detection aims to determine the location of texts from the image. Text recognition is applied to identify the texts and generate a sequence of texts from the text images. Text recognition has been successfully used to extract information from the identity card (ID card) for different languages such as English, Chinese, Vietnamese, Spanish, etc. [3–7]. This study focuses on developing an approach to recognize Vietnamese text and an application to extract information from Vietnamese citizen

identity card (CID card – a.k.a. citizenship card or "Căn Cước Công Dân"). Also, it can be extended to use for the other personal cards.

Much of the work reported in the literature regarding the STR was motivated by the challenges from unstructured texts such as unknown layout, complex background, various view angles, illumination, etc. [1–2]. We should consider the following fundamental research issues to design a realistic system: Feature extraction, classifier, and language model. Conventionally, all these steps in traditional techniques such as hand-crafted feature extraction (HOG descriptor [8], stroke width transform (SWT) [9], maximally stable extremal regions (MSER) [10]) or traditional classifier (conditional random field (CRF) [11], support vector machine (SVM) [12]) are computed or trained separately. As a result, errors at various steps would accumulate during these processes. So, these traditional approaches make the detection and recognition system complicated and inefficient.

The improvements in text detection and recognition have been driven by techniques known as deep learning [13–28] to overcome the drawbacks of the conventional STR system. Yao et al. [13] considered text detection as a semantic segmentation problem. They employed the fully convolutional network (FCN) based on holistically-nested edge detection (HED) to generate global feature maps that include text region information and their relationship. Zhou et al. developed an EAST detector [14] that uses the multi-channel FCN based model for text detection. The detector could generate a text score map and geometry that predict words or text lines of various orientations. Deng et al. applied a PixelLink model [15] that depends on instance segmentation as the text detector. They trained a deep neural network (DNN) for binary classification tasks (text/non-text) and link prediction. The link pixels within text instances were segmented, and the PixelLink located positions of texts from the instance segmentation results without location bounding box regression. Tian et al. [16] developed a deep neural network called a connectionist text proposal network (CTPN) to localized text regions in the image. The CTPN used the VGG16 model as a feature extraction process. They also proposed a vertical anchor mechanism for improving the accuracy of text locations on a fine scale.

In the end-to-end text recognition systems, performing a high accuracy of text recognition needs a large amount of labeled training data. However, text image datasets are often unaligned data. For that reason, we need some extra techniques to map the label sequence to the final text sequence. Recently, Connectionist Temporal Classification (CTC) technique [17–20] and attention mechanism [21–25] have been proposed to train for text recognition systems. Furthermore, convolution neural networks (CNN) [19–20,23] and recurrent neural networks (RNN) – long short-term memory (LSTM) [24–25] have been widely and successfully applied to the text recognition.

Moysset et al. [19] introduced a CRNN model, in which they detected the text lines based on regressions. They proposed applying 2D-LSTM trained with the CTC to identify text lines. Shi et al. [20] developed a novel convolutional recurrent neural network (CRNN) for image-based sequence recognition. In the CRNN, first, they learned a sequential feature representation (a.k.a. feature sequence) from the text image by using the CNN model. Then, they fed the feature sequence to a BLSTM for learning the history and future contextual information. Finally, they used the CTC to get text sequences from the input image. Lee and Osindero [21] introduced a recursive RNN with an attention model for text recognition. They applied recursive CNN for parametrically efficient and effective feature learning. They trained the RNN for learning the character-level language model, which skips using the N-grams model. Finally, the soft-attention model was applied to exploit text image features for training selectively. Cheng et al. [24] proposed a focusing attention network (FAN) that uses a focusing attention mechanism to overcome misalignment problems between feature learning and targets for text images. Shi et al. [25] proposed an end-to-end ASTER model that including a rectification network and a CRNN-based attention network for text recognition. Based on a flexible Thin-Plate Spline transformation, the proposed rectification network

could adaptively an input image into a rectified image. They applied the CRNN-based attention sequence-to-sequence model for predicting text sequences from the rectified image.

Furthermore, Jaderberg et al. [26] employed the CNN and a conditional random field (CRF) together to consider the whole word image as an input for text recognition. They predicted the characters at each location of the output via unary terms of the CRF provided by a CNN. The second CNN then was constructed as a predictor to determine the N-grams model. Yang et al. [27] proposed a simple but effective method, called an adaptive ensemble of deep neural networks (AdaDNNs), for word images. Notably, this method could select and combine classifier components at various iterations by using the Bayesian framework. Liu et al. [28] proposed a novel Fast oriented text spotting (FOST) model for oriented text spotting with a unified framework. They introduced the FCN combining to RoIRotate operator to estimated text bounding boxes. They fed the features of text candidate regions to the RNN and CTC for text recognition.

Recently, some researchers have developed various deep learning approaches to Vietnamese text recognition for constrained forms [6–7,29–30]. Liem et al. [6] proposed an FVI system for Vietnamese identification (ID) card detection and recognition. They first applied RetinaNet [31] to detect the ID card and text lines information. Next, they performed Inception-v3 CNN to extract feature representation from the input text lines. Finally, they used the BLSM combined with spatial attention for text recognition. Viet et al. introduced the corner detection-based Single Shot Multi-box Detector (SSD) with the Mobilenet V2 back-bone for ID card alignment, using the CRNN with soft attention text recognition. Tan and Nam [30] used traditional image processing approaches such as binarization, straight line detection, and heuristic information such as shape, size, the position of text fields for ID cards, and text lines detection. They applied the CRNN based CTC model to recognize the Vietnamese text. In these researches, authors proposed their methods to extract information from the current old forms of Vietnamese identification ID cards, a.k.a. “Chứng Minh Nhân Dân.” This study focuses on developing a robust deep learning approach to recognizing Vietnamese vocabulary words from Vietnam citizen identify a card (CID card, a.k.a. “Căn Cước Công Dân”). This CID is a new form and will replace the current old form of the ID card.

In this paper, we focus on investigating the end-to-end Vietnamese text recognition model and an application to information extraction from the Vietnamese CID card’s front side. We first apply the Mask-RCNN to detect and align the CID card from the background. Next, we introduce two approaches to detect text lines of the CID card’s information field by using traditional image processing techniques compared to the EAST detector. Finally, we proposed a new end-to-end CRNN model based on the combination between the CTC and the Vietnamese text recognition system’s attention. Our model implements the CNN for learning informative representations directly from input text lines image. The feature representation then has fed to the deep BLSTM to learn the history and future contextual features. The proposed model may be directly learned from a sequence of words without detailed annotations using a proposed joint CTC–attention model. The CTC and attention objective functions are jointly trained together. Our main aim in this paper is to explore the effect of the joint CTC–attention model in the CRNN framework. Our model automatically performs alignment between the input text images and recognized words. We use a collected CID card dataset as test cases to verify the performance of the methods.

2 Proposed Approach

We describe the proposed approach in detail. Our approach for extracting information from the front side Vietnamese citizen ID (CID) card includes three main steps, as shown in Fig. 1. We focus on extracting the information from the CID card’s front side because of the essential information present on the front. We first apply the Cropper component as a segmentation method to find the card’s boundary and remove the background. Then, four corners of the card are determined, and then make a perspective transform to

obtain the top-down view. The second component is Text Detector, used to localize and extract the fields of information (i.e., text areas) in the CID card. Finally, the Text Recognizer receives the extracted fields of information as the input and predicts texts.

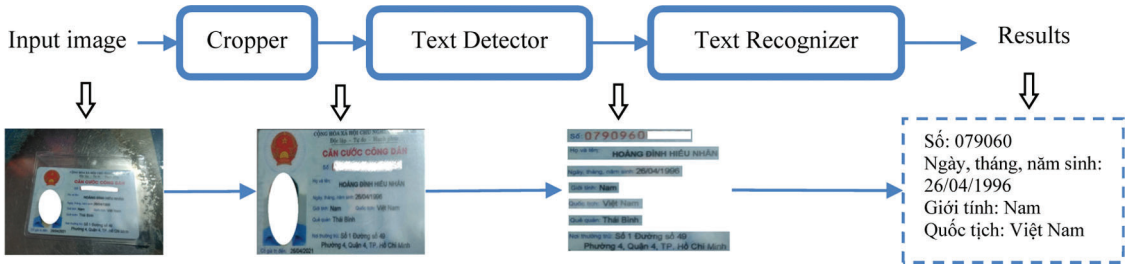

Figure 1: An illustration of the proposed approach for the CID card information extraction

2.1 Cropper

The Cropper is applied to localize the CID card image from the background and convert it into a frontal (top-down) view. In this study, we consider the input images are front side images of CID cards. Traditional approaches based on image processing techniques were proposed to localize and detect card images from the background [3–5,7,30]. Almost these approaches included some main steps as binarization, edge detection, straight line detection, corners detection, and post-processing. However, they only obtain good performance under constraints such as simple background (flat background, text color distinguishing the background of the CID card, background with soft edge, etc.) and no illumination. These approaches were so sensitive to complex backgrounds, illumination effects, blur boundaries, and noise. These techniques have limitations for applying in real cases without constraints even though they were fast and require low memory.

To overcome these problems, we apply a deep learning approach based on the Mask-RCNN model [32] as a binarization method to find the CID card images' boundary. The Mask-RCNN was proven to robust to object detection and segmentation. We assume that no occlusion appears on the card to collect personal information completely and accurately on the card. After detecting the card's boundary in the input image, we apply some image processing techniques to estimate four corners of the card image. Algorithm 1 presents four corners of card image estimation. Finally, we use perspective transform based on methods like *getPerspectiveTransform* and *warpPerspective* in OpenCV [33] to obtain the card image into a rectangular shape resembling a frontal (top-down) view. We resize the output cropped card image before being fed to the Text Detector.

Algorithm 1: Four corners of card image estimation

Input: Boundary of card image (output map from Mask-RCNN).
Output: Four corners of card in image.

Step 1: Use the boundary of card image to find the largest contour by area (outline) representing the card using *findContours* in OpenCV [33].
Step 2: Find rotated rectangle with minimum area using *minAreaRect* in OpenCV [33] that consider the rotation also. Then, determine 04 corners of the rectangle.
Step 3: For each point r_i in the corners of the rectangle (i = 1,…4)
– Find the point c_i on the largest contour so that the distance (Euclidean distance) between r_i and c_i is minimum.
– Then, four points c_1,…, c_4 describe four corners of the card image.

2.2 Text Detector

2.2.1 Proposed Text Detection Method

The Vietnamese CID card image is a fixed layout. The CID card's front side includes some essential information from top to bottom as ID card number, full name, date of birth, gender, nationality, hometown, residence, and expiration date. To take advantage of its shape and layout, we propose a text detection method based on traditional image processing techniques. Our algorithm for text line detection is shown in Algorithm 2.

Given a cropped color card image I, we first resize I to a fixed size of $1200 \times H$, where width is fixed to 1200 pixels to easier to read the text fields information in the card. The height is automatically changed with the same ratio as the width. The aspect ratio is preserved when the same factor scales both width and height. In Step 2, we convert resized color image $Ir(R, G, B)$ into grayscale image J based on the luminosity method as:

$$J = \alpha R + \beta G + \gamma B \tag{1}$$

where the weight factors α, β and γ represent the contribution of red, green, blue colors. We choose the value for these factors are 0.299, 0.587, and 0.114, respectively. After converting to a grayscale image, we need to enhance the background text information before performing binarization from step 3 to step 7.

Given the input's binary image, we extract bounding boxes of connected components from the image based on contour information that is considered text line candidates. We use heuristic information to remove slight noise and unwanted components to obtain the output text line images. Based on the CID card layout, we need to detect eight important information fields such as CID card No., full name, date of birth, gender, nationality, hometown, residence, and expiration date; and remove the other ones. First, based on the location, size, and area of the card image candidates, we can remove unwanted components such as the National Emblem component, identity face photo, headings of the CID card. Next, we determine the bounding box of information fields using fixed thresholds based on their locations and coordinates in the image layout. We next concatenate bounding boxes in the same line using their y-coordinate information to obtain the text line results. Finally, we crop the bounding box of text lines.

Algorithm 2: Proposed text line detection for Vietnamese CID card

Input: Cropped ID card image (output from the Cropper)
Output: Text field images from card image

Step 1: Resize the input image to a size of $1200 \times H$ (fixed width of 1200 pixels).
Step 2: Convert the resized color image to the gray-scale image.
Step 3: Apply Gaussian smoothing with a kernel size of 3×3.
Step 4: Apply a black-hat operator with a kernel size of 15×15.
Step 5: Sobel filter is used to calculate the gradient in the x-direction. The magnitude of gradient then is obtained and normalized into the range of [0,255].
Step 6: Perform closing operator with a kernel size of 31×5.
Step 7: Otsu's method is used for binarization.
Step 8: Find contours and connected component by using the binary image.
Step 9: Extract text line images based on heuristic information such as the characteristic of distance, relative location and length, and coordinate of field information in the image.

From the CID card layout, we consider the residence field (the last information on the front side of the card) usually includes multiple text lines. The second last information that obtains hometown field may consist of one or two lines of text. However, we design the input to our Recognizer component for only a single text line of image-based sequence. So, we need to separate multiple text lines into a single text line. In our experiment, we use the histogram projection technique to separate multiple text lines into cropped regions. Based on the cropped regions' binary components, we use horizontal histogram projection based on a heuristic threshold to separate text in a paragraph. Fig. 2 shows our text line detection method for ID card images.

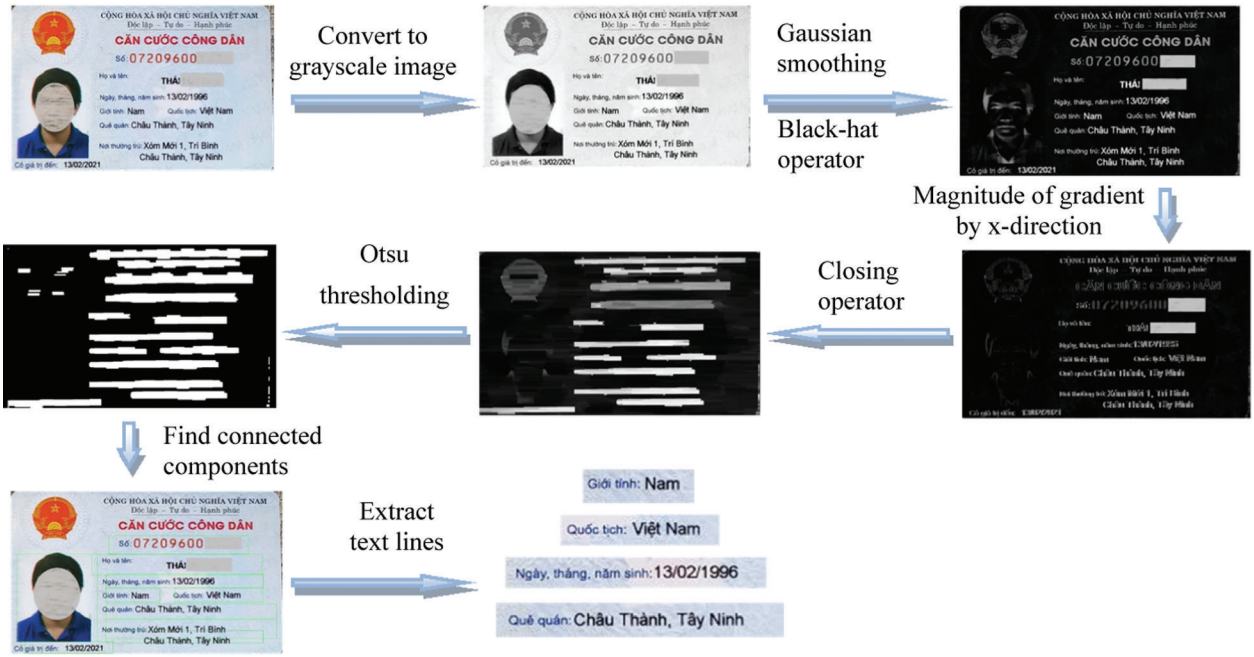

Figure 2: Our Text Detector for extracting text fields

2.2.2 Field of Information Extraction Based on EAST Detector

The EAST detector [14] can detect unstructured texts such as no standard font, no proper row structure, unknown layout, complex background, various view angles, and illumination effects. We can use this detector for detecting non-rectangular regions (such as quadrangles or rotated rectangles) that include the portion of the image having text. Also, this detector can deal with the skewness of the input card image, which traditional image processing approaches cannot easily solve. In our experiment, we first apply the EAST to detect texts in the card images. Then, we normalize the output quadrangles that obtained texts into bounding boxes. We next apply a post-processing technique to bounding group boxes of texts into the same text line using their coordinates as the heuristic information (location, height/width ratio between bounding box size, and the input image size). In our post-processing step, we extend the bounding boxes' size by both height and width to the bounding boxes containing word tokens belonging to each text field. These bounding boxes are sorted depending on their coordinates and then fed to the Recognizer. The extra word tokens values are permanently fixed values for any card images. So, they provide convenient information for the Recognizer and parser procedure. Fig. 3 describes some results of the Text Detector based on the EAST. We consider the EAST detector can deal with the unstructured texts and yield high accuracy, but this method has high computational cost and memory requirements.

Figure 3: Some results of the Text Detector based on the EAST detector

2.3 End-to-End Text Recognition System

We propose an end-to-end CRNN based attention for Vietnamese text recognition by using a combination of convolutional neural network (CNN), recurrent neural network (RNN), and attention mechanism. This combination consists of three components such as convolutional layers, recurrent layers, and transcription layers. We apply this model for image-based sequence recognition based on the idea in [20,25], as shown in Fig. 4. We use the image of text lines as the input to our model. The convolutional layers are first applied to extract feature sequences from the input image. We use the CNN model, followed by an RNN, for predicting the output of each frame of the feature sequence. Finally, we exploit the transcription layer based on the attention mechanism to translate the per-frame predictions into a label sequence (sequence of words). Our model is joint and trained with one loss function.

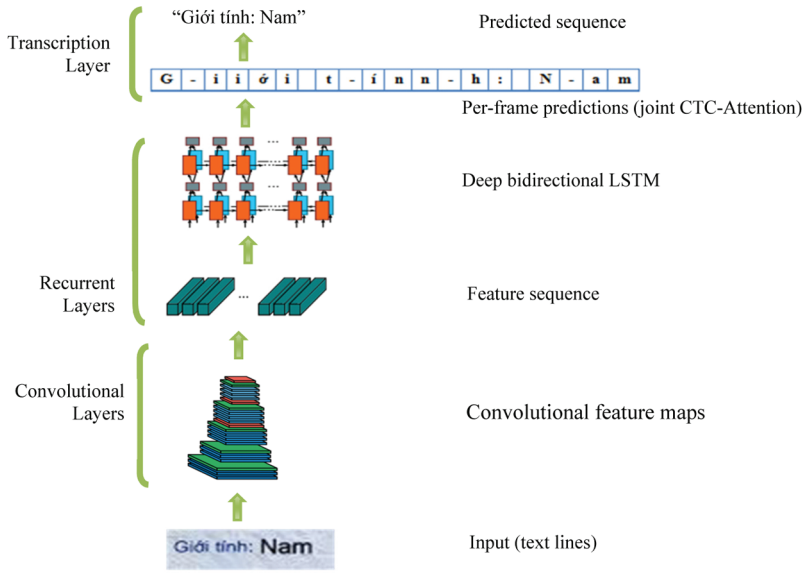

Figure 4: The proposed CRNN model for Vietnamese text recognition

2.3.1 Feature Sequence Extraction

A baseline CNN model which involves convolutional layers, max-pooling layers, and operators such as batch normalization, rectified linear unit (ReLU) activation is used to extract a sequential feature from text line image. This model is designed based on the idea of VGG architecture [34], in which the extracted feature maps are used to create feature vectors, and feature maps of the CNN are used to generate feature vectors of a sequential feature by column (from left to right). We concatenate i-th columns of all the feature maps to

generate the i-th feature vector in the feature sequence. We fix each column's width to a single pixel and use the feature sequence as the input for the RNN.

Because the CNN captures input data's local features at the corresponding position, we consider a rectangle region of the input sequence (termed the receptive field) as a column of the feature maps (corresponding from left to right) [20]. This consideration suits the image-based sequence recognition problem very well because it is considered a sequence of words or characters, which is a significant task for text recognition.

2.3.2 Label Distribution Prediction

We fed the feature sequence from the CNN model to the RNN that used LSTM for learning context information and predicting each frame of the feature sequence's output sequence. The LSTM is well-suited and stable for the image-based sequence of arbitrary length recognition than treating every word independently. A standard LSTM architecture consists of a memory cell storing both past and future information; three gates: an input gate, an output gate, and a forget gate. The input gate discovers extent to which value from input flows into the cell. The output gate controls extent to which the memory cell's value is used to decide the output. Both input and output gates allow the cell to store and retrieve information over long periods. The forget gate is used to discover which value remains in the cell and clear its memory. Therefore, the LSTM can capture long-term dependency adequately, which often occurs in text sequence.

The LSTM is directional, which is used to reflect the input sequence's temporal nature and helps explore the future contextual information. However, a single LSTM only preserves the past contexts because it receives the inputs from the past. In image-based sequence recognition or speech recognition, both past and future information are more practical, and the contextual information can complement each other. So, in this study, we use a combination of two LSTMs, one forward and one backward, into a bidirectional LSTM (BLSTM). The BLSTM discovers both past and future contexts of the input feature sequence [20,25]. Furthermore, we can stack multiple BLSTMs to build a deep BLSTM for image-based sequence recognition [20].

2.3.3 Label Distribution Prediction

The length of the per-frame predictions made by CNN and BLSTM in the previous section may not appropriate to that of the ground-truth label sequence. Therefore, connectionist temporal classification (CTC) [17–20] and attention [21–25] are usually applied to convert the per-frame predictions into a label sequence. These techniques are often to get the label sequence with the highest probability conditioned on the per-frame predictions. The CTC often trains the stacks of LSTM to maximize the probability of the label sequences in a training set. Using CTC, we can obtain both alignment and recognition, and CTC can map the input sequence to the label sequence without segmentation and alignment between the input and the output. In this study, we propose the attention-based encoder-decoder approach for image-based sequence recognition in Vietnamese. It can perform the alignment between the input image and the output label sequence (recognized strings).

We assume that the probability is defined for model output L-length character sequence C (called label sequence) conditioned on the input per-frame predictions, $y = \{y_1, \ldots, y_T\}$ that is output from deep BLSTM, where T is the sequence length. Given the input, per-frame predictions, $y = \{y_1, \ldots, y_T\}$ is the output from deep BLSTM, and the attention mechanism is used and directly estimate the posterior probability of the L-length character sequence C (called label sequence), $p(C|y)$. Compared to the CTC approach, the attention mechanism may not make any conditional independence assumptions. The posterior probability $p(C|y)$ is computed based on the chain rules as follows:

$$p(C|y) = \prod_{l} p(c_l|c_l, c_2, \ldots, c_{l-1}, y), \tag{2}$$

where $p(c_l|c_l, c_2, \ldots, c_{l-1}, y)$ is estimated by the attention-based objective function from [35], as follows:

$$h_t = Encoder(y), \tag{3}$$

$$a_{lt} = Attention(\{a_{l-1}\}_t, q_{l-1}, h_t), \tag{4}$$

$$r_l = \sum_t a_{lt}\, h_t, \tag{5}$$

$$p(c_l|c_l, c_2, \ldots, c_{l-1}, y) = Decoder(r_l, q_{l-1}c_{l-1}) \tag{6}$$

where h_t denotes a hidden vector that is converted from input feature y via the encoder. The attention in Eq. (4) uses a content-based mechanism with convolutional features [35]. The attention weight a_{lt} represents a soft alignment of the hidden vector h_t for each output c_l. It is used to compute the context vector r_l for the decoder. The decoder generates the output label sequence from the encoded sequence. It considers the previous c_{l-1} and decoder hidden state q_{l-1} as input values and provides the cumulative context of the decoder's predictions into the next prediction [35]. In this study, the BLSM is used as a recurrent layer in the encoder and decoder neural network.

2.4 Joint CTC and Attention Training

Given a training dataset, $D = \{d_i, C_i\}_i$, where d_i is the training text line image and C_i is the ground truth label sequence. The CTC loss function that minimizes the negative log-likelihood of the conditional probability of ground truth is shown as follows [20]:

$$\mathbb{C} = -\sum_{d_i, C_i \in D} log\ p(C_i|y_i), \tag{7}$$

where y_i is the output sequence extracted from the BLSTM and convolutional layer from d_i. The loss function of attention is calculated based on cross entropy as:

$$\text{Å} = -\sum\nolimits_{d_i, C_i \in D} \sum\nolimits_{c_j \in C_i, t_j \in y_i} p(c_j) log(q(t_j)), \tag{8}$$

where $p(c_j)$ is probability of the output j-th element of the label sequence C_i, and $q(t_j)$ is a probability of the j-th element of the output sequence y_i of attention corresponding to the input d_i.

In our research, we propose to apply the attention mechanism to the CRNN framework to work more efficiently on long image sequences. However, this mechanism leads to time-consuming when the data-driven attention is only used for estimating the label sequences in long image sequences. To speed up the label sequence estimation process and reduce irregular alignments during training and inference, we consider the length of output label sequence from CTC as additional information to attention-based decoder predictions to make the final sequence without any manual effort. Besides, to make independence between two CTC outputs and attention, we use the first BLSTM's output in the RNN layer as the encoder's input. So, an additional encoder-decoder RNN with attention helps improve the performance of the text recognition system. Finally, the loss function of our joint CTC-Attention network architecture is combined as follows:

$$\mathcal{L} = \lambda \mathbb{C} + (1 - \lambda) \mathring{\mathrm{A}} \tag{9}$$

where $0 \leq \lambda \leq 1$. In the experiment, we set $\lambda = 0.5$.

To train the network, we use stochastic gradient descent (SGD) via the back-propagation algorithm. The error differentials are back-propagated using the forward-backward algorithm in the transcription layer, and the Back-propagation Through Time is applied to calculate the error differentials in the recurrent layer [20]. Also, we use the ADADELTA algorithm [20] to compute per-dimension learning rates for optimization.

3 Experiments and Results

3.1 Dataset

We carried out all experiments using the dataset of Vietnamese citizen identity card (CID) images. These real front side card images of people volunteers were taken from the camera, named the CID dataset. Because the CID dataset was personal data that requires private and confidential data, collecting the real CID images is quite tricky, and the number of images is minimal. We collected a total of 120 real front side card images and their corresponding ground-truth words. We only used this dataset for research purposes. The dataset is taken from many people from three main regions: North, Central, and South Vietnam with significant variations, such as contrast, background, and illumination. The dataset's size is relatively small for the current standard end-to-end text recognition system, which was a challenge to our system. Fig. 5 shows some CID images in our dataset.

Figure 5: Examples of CID card images in our dataset

In this study, we only used our real CID dataset for testing. We used synthetic data and augmentation data to train our framework that includes three components: Cropper, Text Detector, and Text Recognizer. To train the Cropper based on the Mask-RCNN, we generated synthetic un-cropped CID images and made augmentation data. The artificially created data consisted of pasting frontal cropped card images onto various backgrounds and then performing random transformations, such as rotation, noise, and blur. There are 1150 synthetic un-cropped CID images used as the training set, and the test dataset only uses 120 real-world CID datasets.

To evaluate the performances of the Cropper, Text Detector, and Text Recognizer components separately, we first prepared the test dataset for Text Detector by manually cropping our real-world input images and refining them to a frontal (top-down) view. To evaluate the EAST detector's performance, we did not perform any change or refinement to the EAST framework. Instead, we used the network as a black box and did not perform retraining. We used our real cropped dataset for testing.

There are not many Vietnamese CID images used for research purposes, and most Vietnamese CID datasets are not shared with the public due to security reasons. Under these circumstances, researchers still have minimal access to extensive real Vietnamese CID data. To deal with this problem, we generated a synthetic dataset for training our models. The CID synthetic dataset contains 73090 text line images that obtain the CID card's information field and their corresponding ground-truth words. Text lines in the CID synthetic dataset are created using data from the Vietnamese administrative divisions, including full

Vietnamese name, date of birth, etc. The text line images are generated by a synthetic text engine and are highly realistic. Fig. 6 presents our synthetic CID text line images dataset.

Figure 6: Our synthetic text line images for training

For text recognition, we only used the synthetic CID dataset as training data. Thus, for the synthetic dataset, 70044 text line images were selected as the training set, and another 3046 text line images were used as a validation set. The test dataset only used our real test CID dataset without any fine-tuning on training data. Finally, there were 1065 real cropped text line images used to evaluate text recognition performance.

Besides, Vietnamese are so complex because Vietnamese commonly has six tones: on unmarked and five marked ones: grave accent, hook above, tilde, acute accent, and a dot below. There are several phonetic variations in the tones among dialects. These diacritical marks are small and not easy to recognize. Some Vietnamese words have similar character components, but different diacritical marks generate six other words with a different meaning such as "ma" (meaning: funeral), "má" (mother), "mà" (but), "mã" (horse), "mả" (grave), "mạ" (rice seedling). Hence, the recognition errors can occur due to missing diacritical marks, wrong letters with similar shapes, etc., such as — huu – hưu, nhiêu– nhiều, mang–mong.

Furthermore, the text line images usually have a different background, font size, and length. These differences are considerable evidence that building good-performance Vietnamese text recognition is quite a difficult task.

3.2 Analysis of the Experiments

We used the Mask-RCNN model architecture in the Cropper component with the pre-trained model on the COCO 2017 dataset [36]. The EAST detector model architecture was adapted from the framework using the ResNet-50 model with pre-trained on the ICDAR 2015 dataset. We implemented the Mask-RCNN and EAST detector models using the Tensorflow framework. The input text line images first were resized to a height of 32 pixels before feeding to the CNN component in the CRNN model. The decoding target vocabulary includes all the 108 characters (93 characters that includes 89 characters in Vietnamese alphabet with tones and 04 characters ("f", "j", "q", "w" and "z"), 10 digits and 05 special tokens {',' ; '.' ; '-' ; '/'; ':' }) that occurred in the transcriptions.

We used the VGG-Net for the CNN layers. We designed the network to make feature maps with a larger width and a longer feature sequence. We used the batch normalization in the CNN layers to limit the covariate shift by fixing layer inputs' means and variances. The CNN included 07 convolutional ReLU layers and 04 max-pooling layers. The RNN layer included 02 layers of BLSTM and a linear layer. The BLSTM adopted a time-step of 512 (length of the input sequence), and the number of features per time-step (number of hidden units) was 256. We introduced the attention-based encoder-decoder to improve the performance for a long sequence. We implemented the text recognition model using the Pytorch framework.

The model was trained on a workstation with Intel(R) Xeon(R) Silver 4210 CPU @ 2.20 GHz, 64 GB RAM, and NVIDIA Quadro K6000 GPU. The training process took less than 05 days to reach convergence. The word error rate (WER) was used as the measurement to evaluate the text recognition system's accuracy.

3.3 Results and Comparisons

3.3.1 Experimental Results on the Cropper

We compared the Mask-RCNN model's performance with various backbones such as Resnet50, Resnet101, InceptionResnet V2. Tab. 1 shows the performance and computation cost (in average time) for the Cropper.

Table 1: The accuracy and computation cost of the Cropper component for the CID image alignment

The Cropper	Accuracy (%)	Computation cost (second)
Mask-RCNN Resnet 50	96.13%	**0.50 s**
Mask-RCNN Resnet 101	95.92%	0.52 s
Mask-RCNN InceptionResnet V2	**98.85%**	0.59 s

Experimental analysis revealed that we achieved the best accuracy through the Mask-RCNN with InceptionResnet V2 backbone being 98.85%. However, there were certain disadvantages with computation cost. Fig. 7 shows the results of the Mask-RCNN for four corner detection and the CID alignment. This figure clearly shows that the proposed method had achieved high performance. Our Cropper method had demonstrated promise for enhancing the performance of the CID image alignment. The proposed method was general to be applied to other documents, paper sheets, personal cards, invoice bills, etc., for document alignment.

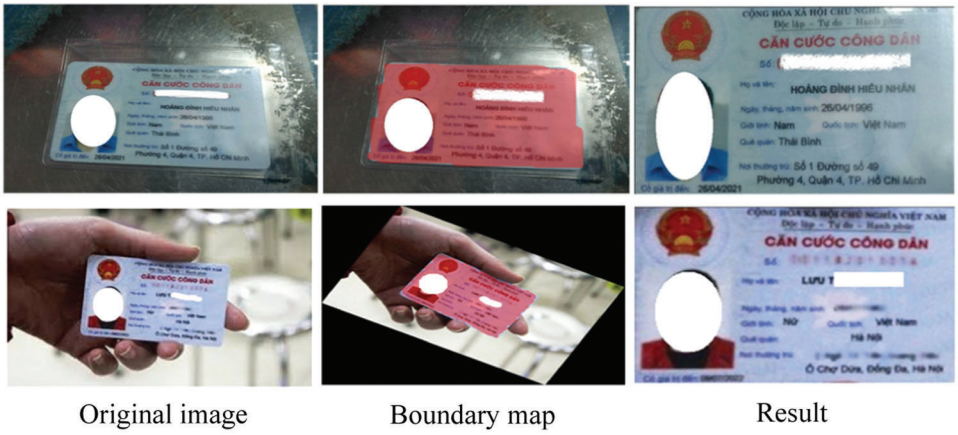

Figure 7: Results of the Mask-RCNN for the CID alignment

3.3.2 Experimental Results on Text Detector

We investigated the effectiveness of the proposed text detector approach compared to the EAST detector for text line image extraction. Each CID image includes eight information fields. We evaluated the text detector results based on the number of extracted text line images compared to the ground-truth based on the IoU of 0.4. Tab. 2 shows the text detector's performance based on the proposed traditional approach and the EAST detector.

Table 2: Comparison of the text detector using traditional image processing and EAST detector approaches on our CID image dataset

Methods	Precision (%)	Recall (%)	F1-score (%)	Computation cost
Traditional image processing	0.886	0.890	0.888	**0.027 s**
EAST detector	**0.946**	**0.945**	**0.945**	0.402 s

Our proposed approach and EAST detector's accuracy was also greater than 88% in precision, recall, and F1-score. Especially, our approach could obtain 88.80% in F1-score. The EAST detector performed 94.50% in F1-score. Experimental analysis indicated that the EAST detector yields higher accuracy than the traditional approach. In the conventional approach, we used a fixed threshold based on the heuristic information such as shape, size, the position of the text field in the CID image for text line extraction. It could lead to false alarms or miss detection because of detection errors.

Furthermore, we also measured computation time (in average) between these two approaches. From Tab. 2, the conventional method performed faster than the EAST approach when it only took 0.027 s for extracting text lines from a single CID image compared to the EAST detector took 0.402 seconds for each single CID image. The traditional approach was lower computation cost than the EAST detector and did not require GPU computing. Also, we analyzed and evaluated the performance of text detectors in each information field. Tab. 3 shows the results of text detectors in each information field. This table clearly shows that our two approaches obtained very high accuracy for five information fields: ID number, full name, date of birth, gender, and nationality. Most errors occurred while processing the last three fields as hometown, residence, and expiry date. The EAST detector obtained very high accuracy for seven fields except for the residence field. These errors were caused by the small text size in the last region ("expiry date"), and this text lay near the margin of the CID card. Also, hometown, residence regions sometimes have two text lines because they usually include many words. The locations of text lines in these two regions often deviated from the CID layout. Due to the text line's deviation, each horizontal text line in a field could overlap with other text lines from another field, affecting the text detection approach's performance. The conventional approach could obtain detection errors due to these problems because it used a pre-define threshold based on the location, shape, or size of text lines. In our cases, some images could be affected by illumination, such as low-light conditions, low contrast, etc., reducing our text detection system's accuracy. Fig. 8 shows some text line images from hometown and residence fields affected by illumination and deviation from the CID layout.

Table 3: Results of text detector in each information fields

Information fields	Accuracy (%)	
	Traditional approach	EAST detector
ID number	**97.57**	**99.39**
Full Name	**93.93**	**96.36**
Date of birth	**96.96**	**96.96**
Gender	**96.36**	**98.18**
Nationality	**95.57**	**98.18**
Hometown	80.39	**96.96**
Residence	84.24	87.87
Expiry date	86.67	**93.93**

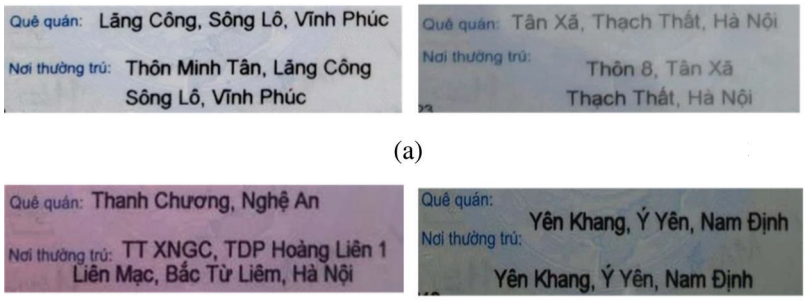

Figure 8: Some text lines images affected by illumination and deviation from the CID layout

3.3.3 Experimental Results on Text Recognition

We first explored the effects of the CRNN based attention mechanism on manually cropped text line images from the CID card images. Tab. 4 shows the performance comparison in WER of the CRNN model based on the CTC and attention. In our experiments, we referred to the CRNN+CTC and CRNN +Attention as baseline methods. Through Tab. 4, the proposed method obtained a good performance compared to the baseline methods. We realized that the CRNN with CTC and attention mechanism achieved better accuracy. Specifically, the WER value was 4.28%. The WER rate using the proposed approach attained 6.77% and 4.28% for CRNN+Attention and joint CRNN+CTC+Attention, respectively. Our CRNN+joint CTC-Attention performance outperformed the baseline methods on the CID dataset.

Table 4: Performance comparison of CRNN with and without attention mechanism for Vietnamese text recognition using manually cropped information fields

Methods	WER (%)
CRNN+CTC [20,25]	5.68
CRNN+Attention [25]	6.77
CRNN+ joint CTC-Attention (proposed)	**4.28**

In the end-to-end text recognition system, the CRNN+CTC required a large enough dataset to get better performance. With the limitation of training data, the additional encoder-decoder BLSTM with attention significantly improved the text recognition system's performance. Furthermore, the attention mechanism led our model to work more efficiently on long input sentences. These results proved that the CRNN with attention mechanism led to greater robustness, significantly improved the Vietnamese text recognition system's performance, and outperformed the standard approach.

We attributed these mistakes to the process of output sequences. The errors often occurred due to unrecognized small characters, confusion between characters, or missing Vietnamese diacritical marks in the certain font on the CID card such as "6" - "8", "6" - "0", "." - ",", "D" - "Đ" and "a"-"â", etc. Besides, in the certain font on the CID card image, the Vietnamese diacritical marks were small. Low contrast or illumination in the input text image led to difficulty recognizing. During the processing at the CNN layers, we might remove the diacritical marks, so the prediction will often lack information such as "Văn" – "Vân," "Thi" – "Thị," "Ấp" – "Âp," "TRẦN" – "TRÂN" – "TRẤN," "Lăng" – "Lãng." These terms were the most challenging for Vietnamese text recognition tended to cause more distortion and

reduced the accuracy of the Vietnamese text recognition system. In our study, we did not use any post-processing method for spell or text correction.

We also evaluated our text recognition performance through eight information fields' accuracy, as shown in Tab. 5. We manually cropped information fields before feeding them to our CRNN based joint CTC-attention model. The table indicates that our proposed method yielded good performance for Vietnamese CID card recognition. The proposed method obtained very high accuracy for such information fields as date of birth, name, gender, expiry date, and ID number. Some misclassification errors while processing hometown and residence fields. These fields were often represented by a long text sequence image with small Vietnamese diacritical marks. We also compared the performance of a combination between the text detector and recognizer. The EAST detector's output text line images were fed to the CRNN based joint CTC-attention model and evaluated. The recognition performance of these systems is shown in Tab. 6.

Table 5: Performance of text recognition system for each information fields in the CID card

Information Field	WER (%)
ID number	**4.66**
Name	**0.79**
Date of birth	**0.00**
Gender	**0.22**
Nationality	**0.00**
Hometown	6.22
Residence	12.64
Expiry date	**0.04**

Table 6: Comparison of the CRNN+CTC+Attention using EAST text detector approach on our CID dataset

Methods	WER (%)	Computation cost (in average)
EAST+CRNN+CTC+Attention	5.38	0.4239 s
Manual cropped+CRNN+CTC+Attention	**4.28**	**0.0219 s**

From this table, we realized that the CRNN+CTC+Attention using the EAST detector obtained high performance with the WER of 5.38%. The table results indicated that the accuracy of the CRNN+CTC +Attention using the EAST detector was close to the results of the CRNN+CTC+Attention using manually cropped text line images. The recognition system, with a combination to EAST detector, was proven to performed high accuracy. However, the EAST detector required high costs in terms of computational complexity than the traditional approaches. Furthermore, we measured the computation time of our CRNN+CTC+Attention model. From Tab. 6, our model only took 0.0219 seconds for recognizing a single text line image.

4 Conclusions

We proposed an improved Vietnamese text recognition technique by investigating a combination between the CTC and attention training for the CRNN model. We considered the effect of the attention

model on enhanced Vietnamese text recognition. The experimental results prove that the proposed approach could significantly improve the recognition system's accuracy and outperformed the standard CRNN method. Our approach was effective and robust under noise environments. Our model also maintained its generality when applying to other languages. In the future, we will consider an exploration of the effects of different types of CNN and attention methods to compare their effectiveness in Vietnamese text recognition systems.

Funding Statement: This research was supported by Sai Gon University under Fund (Grant No. TĐ 2020-11).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Lin, P. Yang and F. Zhang, "Review of scene text detection and recognition," *Archives of Computational Methods in Engineering*, vol. 27, no. 2, pp. 433–454, 2020.

[2] X. Chen, L. Jin, Y. Zhu, C. Luo and T. Wang, "Text recognition in the wild: A survey," *arXiv preprint arXiv:2005.03492*, 2020.

[3] J. Lladós, F. Lumbreras, V. Chapaprieta and J. Queralt, "ICAR: Identity card automatic reader," in *Proc. of Sixth Int. Conf. on Document Analysis and Recognition*, pp. 470–474, 2001.

[4] H. Lee and N. Kwak, "Character recognition for the machine reader zone of electronic identity cards," in *IEEE Int. Conf. on Image Processing (ICIP)*, pp. 387–339, 2015.

[5] M. Ryan and N. Hanafiah, "An examination of character recognition on ID card using template matching approach," *Procedia Computer Science*, vol. 59, pp. 520–529, 2015.

[6] H. D. Liem, N. D. Minh, N. B. Trung, H. T. Duc, P. H. Hiep *et al.,* "FVI: An end-to-end Vietnamese identification card detection and recognition in images," in *5th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 338–340, 2018.

[7] N. T. T. Tan and N. T. Khanh, "A method for segmentation of Vietnamese identification card text fields," *Advanced Computer Science and Applications*, vol. 10, pp. 415–421, 2019.

[8] K. Wang, B. Babenko and S. Belongie, "End-to-end scene text recognition," in *Int. Conf. on Computer Vision*, pp. 1457–1464, 2011.

[9] B. Epshtein, E. Ofek and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2963–2970, 2010.

[10] L. Guan and J. Chu, "Natural scene text detection based on SWT, MSER and candidate classification," in *2nd Int. Conf. on Image, Vision and Computing (ICIVC)*, pp. 26–30, 2017.

[11] Y. F. Pan, X. Hou and C. L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 800–813, 2010.

[12] J. Almazán, A. Gordo, A. Fornés and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[13] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou *et al.,* "Scene text detection via holistic, multi-channel prediction," *arXiv preprint arXiv:1606.09002*, 2016.

[14] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou *et al.,* "EAST: An efficient and accurate scene text detector," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651, 2017.

[15] D. Deng, H. Liu, X. Li and D. Cai, "PixelLink: Detecting scene text via instance segmentation," in *Proc. of the 32nd AAAI Conf. Artificial Intelligence*, New Orleans, LA, USA, pp. 6773–6780, 2018.

[16] Z. Tian, W. Huang, T. He, P. He and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," In: B. Leibe, J. Matas, N. Sebe, M. Welling (eds.), *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, Springer, Cham, vol. 9912, pp. 56–72, 2016.

[17] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML'06: Proceedings of the 23rd international conference on Machine learning, June 2006*, pp. 369–376, 2006.

[18] A. Ray, S. Rajeswar and S. Chaudhury, "Text recognition using deep BLSTM networks," in *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, pp. 1–6, 2015.

[19] B. Moysset, C. Kermorvant and C. Wolf, "Full-page text recognition: Learning where to start and when to stop," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, pp. 871–876, 2017.

[20] B. Shi, X. Bai and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[21] C. Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for OCR in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2231–2239, 2016.

[22] F. Bai, Z. Cheng, Y. Niu, S. Pu and S. Zhou, "Edit probability for scene text recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1508–1516, 2018.

[23] C. Bartz, H. Yang and C. Meinel, "STN-OCR: A single neural network for text detection and text recognition," *arXiv preprint arXiv:1707.08831*, 2017.

[24] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu *et al.,* "Focusing attention: Towards accurate text recognition in natural images," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5076–5084, 2017.

[25] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao *et al.,* "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2018.

[26] M. Jaderberg, K. Simonyan, A. Vedaldi and A. Zisserman, "Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*, 2014.

[27] C. Yang, X. C. Yin, Z. Li, J. Wu, C. Guo *et al.,* "AdaDNNs: Adaptive ensemble of deep neural networks for scene text recognition. *arXiv preprint arXiv:1710.03425*, 2017.

[28] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao *et al.,* "FOTS: Fast oriented text spotting with a unified network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5676–5685, 2018.

[29] H. T. Viet, Q. H. Dang and T. A. Vu, "A robust end-to-end information extraction system for Vietnamese identity cards," in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 483–488, 2019.

[30] N. T. T. Tan and N. H. Nam, "An efficient method for automatic recognizing text fields on identification card," *VNU Journal of Science: Mathematics-Physics*, vol. 36, no. 1, 2020.

[31] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[32] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[33] The Open CV Reference Manual, Release 2.4.13.7. [Online]. Available: https://docs.opencv.org/2.4/opencv2refman.pdf, 2019.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] T. Hori, S. Watanabe, Y. Watanabe and W. Chan, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *Proc. Interspeech*, pp. 949–953, 2017.

[36] Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4278–4284, 2017.