Text Region Localization and Recognition for ID Card Identification using Deep Learning Approaches

Saw Zay Maung Maung
University of Computer Studies, Yangon
Yangon, Myanmar
sawzaymaungmaung@ucsy.edu.mm

Nyein Aye
University of Computer Studies, Yangon
Yangon, Myanmar
nyeinaye@gmail.com

***Abstract*—Deep Learning approaches are currently successful in the application areas of object detection and recognition. In Optical Character Recognition (OCR) applications, there are still challenges in the problems of text region localization and segmentation. Various conventional methods that used in OCR systems still couldn't get high accuracy and still to do to be more effective for real time OCR applications. This paper presented recognition and identification of the specified identity document (ID) to solve those problems by applying deep learning techniques. For the process of text region extraction, the Regional Proposal Networks (RPN) such as Faster R-CNN is used to identify which areas are the most probable of text in the ID card. For text sequence learning, Recurrent Neural Network (RNN) learns timestamp-based text sequences from images to avoid segmentation problems and recognize a given text. According to the experimental results, the deep learning techniques give a satisfied accuracy and loss rates in text region identification and recognition for ID cards.**

Keywords— Optical Character Recognition, Regional Proposal Networks, Recurrent Neural Network, sequence learning

I. INTRODUCTION

One of the most significant application areas in the computer vision and image processing fields is Optical Character Recognition (OCR). In general, OCR has three processes to recognize optical characters in an image: preprocessing, text region segmentation (extraction), and character recognition. To get accurate information from the text images, the result of each process is very important for its next processes.

The texture-based and region-based are two commonly used techniques for text region extraction. The process of the texture-based techniques is based on the top-down approach. Thus, the texture features are extracted first and then text regions are found in this process. Some of the Texture-based techniques are Fast Fourier Transform(FFT), Gabor filters, spatial variance, and Wavelet. Contrarily, the opposite of Texture-based techniques, region-based techniques use the bottom-up approach. Hence, in this technique, segmentation is performed in small regions and then grouped by similar potential text regions.

One of the region-based technique is Regional Proposal Network (RPN) that predict object scores and object bounds simultaneously at each position by using fully convolutional network. It has a classifier and a repressor. For generating the high-quality region proposals, RPN is trained end-to-end deep learning network model. Currently, RPN is used in a deep learning method and Faster R-CNN is one of the types of RPN for real time objects detection and recognition. It can also be used effectively in extracting text regions.

For recognizing the text from document images, conventional OCR system is mostly based on artificial neural network or template matching. However, these two methods produce less accuracy. In conventional methods, the text images are segmented into character level or word level projections before the recognition process. If the process of segmentation is not accurate, then text recognition will not get high accuracy. Currently, deep learning-based sequence learning is used in OCR to avoid segmentation problems. Long Short-Term Memory Network (LSTM) is capable of learning long-term dependencies and provides an ability for segmentation-free recognition that take raw pixel data as input [1-3].

The main objective of this paper is to enhance the process of academic tasks by automatic recognizing ID cards in universities using image processing and computer vision techniques. In university, student ID cards are essential for processing academic tasks such as filling data when students borrow items from the library or when they take examination. All of these tasks need to view student ID cards, and then retype and fill data into the database for recording the student data. According to the objective, this paper used RPN for extracting the text regions in recognizing the ID cards in order to avoid manual filling data. Moreover, deep learning-based sequence learning is used in recognizing each character for getting high accuracy.

The rest of this paper is organized as follows: section 2 lists the related works of this paper. Section 3 describes the background theory of Regional Proposal Network, convolutional neural network for feature extraction, and sequence learning for text region recognition. And section 4 introduces the design of process flow for this recognition system. Section 5 describes the experiment of the proposed system and the evaluation of the results. Section 6 describes the conclusion and future works.

II. RELATED WORKS

With advanced technology, Object Detection methods based deep learning approaches are used in the application areas of OCR. The application areas are text spotting which involves in recognizing image areas of text, such as a wall plaque or a sign by using recurrent neural network such as bi-directional LSTM [4-5].

A Mongolian OCR system is proposed in [6] to recognize the Mongolian words directly without segmentation. The size of vocabulary in Traditional Mongolian language is very large. This system used sequence to sequence mapping to be easier in training. Moreover, it also intends to overcome the out-of-vocabulary for getting the relationship between letters and glyphs.

Yousfi et al. proposed text recognition methods without using segmentation or pre-processing [7]. As a replacement for using hand-crafted features, they learned features with Convolution Neural Network (ConvNet) and Deep Auto-Encoders. Then, they used multi-scale sliding window scheme, and connectionist recurrent approach for feature extraction, and text decoding respectively. To classify an accurate transcription of the input image, they trained the associated sequence of features.

A Chinese image text recognition system is developed by using an image text that includes a line of character as input, performing the recognition on the whole image text level, and producing a recognized text sequence [8]. They applied the BLSTM-CTC (bidirectional LSTM-Connectionist Temporal Classification) scheme for developing this system and training a neural network by collecting over 2 million news titles. They mapped input sequence into a vector of a fixed dimensionality and used CTC for decoding.

The identification system of ID card number is also presented by Zhu et al. [9]. They used a Convolutional Neural Network model, LeNet-5 network model, for identifying the ID card by selecting the appropriate channel component of color distribution because this model gives the high effectiveness in the character recognition application. The projection method is also used to extract every single character.

Mollah et al. [10] used rule-based classification to identify only the text regions of the card and region growing approach to identify connected component of every character in designing the Business Card Reader (BCR) for mobile devices.

According to the description of the related papers, OCR systems for any language developed by using conventional methods have segmentation problems. Thus, the segmentation avoidance approaches are needed to implement these systems for gaining high accuracy. Deep learning-based sequence learning approach is commonly used for tackling these problems. Specifically, LSTM has the ability to learn long-term dependencies and recognize the raw pixel input data without segmentation. In addition, the process of text extraction is also important in the text recognition process. Thus, this paper used deep learning-based sequence learning approaches for automatically recognizing the student ID cards with high accuracy. Moreover, it also used RPN to extract the text regions in the process of recognizing ID card for the avoidance of filling data manually.

III. Background Theory

Generally, developing an OCR application has three steps such as preprocessing (binarization, noise removing, etc.,), text region recognition and text recognition. To get improvement on accuracy, various methods (conventional or deep learning) are used to solve OCR problems through these three steps. In this system, deep learning approaches are used to localize text regions and text recognition. Thus, this section described the two techniques: region proposal network and sequence learning for localizing text regions and text recognition respectively.

A. *Region Proposal Network*

One of the Deep learning-based object detection approaches, Regional Proposal Network (RPN), is used for the detection text region in an image. It uses any size of an image as input and produces a set of rectangular object proposals with its objectness score as output. The region proposals are generated by sliding a small network over the convolutional(conv) feature map output of the last shared conv layer. Each sliding window is mapped to a lower-dimensional vector and then fed it into two siblings fully-connected layers: a box-classification layer(cls) and a box-regression layer (reg).

Faster R-CNN is used by RPNs to train end-to-end region proposals-based object detection as shown in Fig. 1 [11]. Faster R-CNN uses a convolutional neural network. It is composed of three parts – Convolutional layers for Feature Extraction, Regional Proposal Network, and Class and Bounding boxes predictions. In the first part, the appropriate features from the image are extracted by filtering with convolutional layers. In the second part, RPN slides on the last feature map of the convolution layers and predicts whether there is an object or not. Moreover, it also predicts the locations of bounding box for the objects. In the third part, fully connected neural networks are used to takes the regions proposed by the RPN as an input and predicts Bounding boxes (Regression) and object class (classification).

An anchor is a box and plays an important role in Faster R-CNN. In Faster R-CNN's default configuration, each position of an image has 9 anchors [12] and an image which has 9 anchors at the position (320, 320) with the size of (600, 800).

The training loss used in RPN is also a multi-task loss. The loss function can be defined by equation (1) [13].

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) \quad (1)$$

In the equation (1), i is the indexes of the anchor points in the mini-batch images. $L_{cls}(p_i,p_i^*)$ is log-based classification loss over two classes (object or not object). The p_i is an output score from the classification branch for anchor i, and p_i^* is the ground truth label (0 or 1). The regression loss $L_{reg}(t_i,t_i^*)$ is activated when the anchor actually contains an object i.e., the ground truth p_i^* is 1. The output prediction of the regression layer t_i consists of 4 variables $[t_x,t_y,t_w,t_h]$. The regression target t_i^* can be calculated by equation (2).

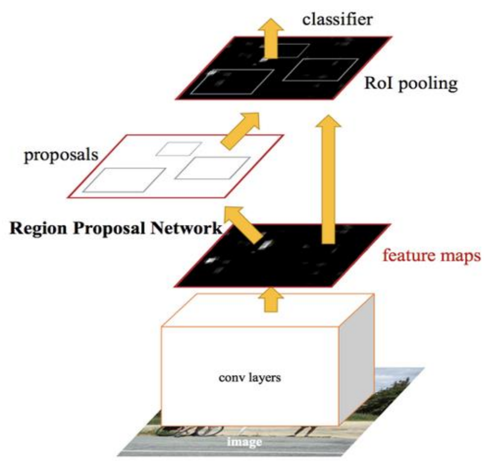

Fig. 1. Design of a faster R-CNN [11]

$$t_x^* = \frac{x^* - x_a}{w_a}, t_y^* = \frac{y^* - y_a}{h_a}, \tag{2}$$
$$t_w^* = log\left(\frac{w^*}{w_a}\right), t_h^* = log\left(\frac{h^*}{h_a}\right)$$

In equation (2), x, y, w , and h correspond to the (x, y) coordinates of the top-left coordinate and the height(h) and width(w) of the object bounding box. x^*, x_a stand for the coordinates of the anchor box and its corresponding ground truth bounding box.

B. *Convolutional Neural Network for feature extraction*

Convolutional Neural Network (CNN) is a deep learning model and basically used in feature extraction and image recognition. In CNN, many different types of layers (Convolution layer, Rectified Linear Unit, Pooling, Dropout, fully connected layers) are stacked sequentially to process image pixels. It uses "n x n" convolution filter to slide over image regions and extracts feature maps for further object recognition tasks. For reducing dimensionality with losing a few of features, pooling (primarily max_pooling) layers, activation function layers, and convolution layers are included in it. Various CNNs architectures (LeNet, AlexNet, VGGNet, etc.,.) are available in building deep learning-based applications. Most End-to-End deep learning includes combination of CNN and RNN to provide a framework for real time applications. In this system, CNN is learned feature maps automatically in a given image for text recognition.

C. *Sequencing Learning for Text Recognition*

For Sequence Learning to avoid segmentation problems in an OCR application, Recurrent Neural Network learns text sequences from timestamp to timestamp. The function of loops in Recurrent Neural Networks is the network activations from a previous time step are fed as inputs to the network for predicting effectively at the current time step. LSTM is one of the types of Recurrent Neural Network and has the ability of learning long-term dependencies. LSTM networks provide an ability for segmentation-free recognition that take raw pixel data as input. LSTMs also have this chain like structure as shown in Fig. 2, but the repeating module has a different structure: LSTM Weights – (Input, Output, Internal State), LSTM Gates – (Forget gate, Input gate, Output gate).

The learning equations of LSTM network are shown in equation (3), (4), (5), (6), and (7).

The output of LSTM network for sequence learning is the probabilities matrix for the distribution of all class labels. For the classification of temporal tasks and sequence labeling problems, the alignment between the inputs and outputs is unknown, CTC loss function measures the distance between softmax activation and ground truth labels.

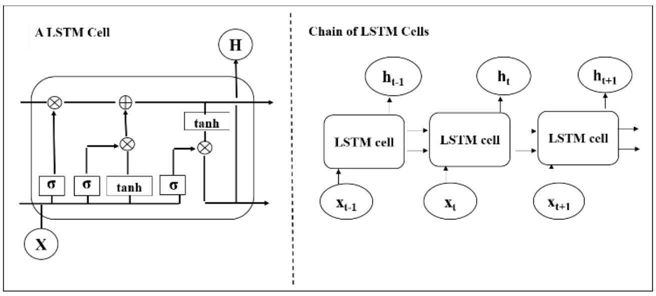

Fig. 2. LSTM network

$$o_t = \sigma(W_0[h_{t-1}, x_t] + b_0) \tag{3}$$

$$h_t = o_t * \tanh(C_t) \tag{4}$$

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{5}$$

$$\bar{C_t} = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{6}$$

$$\bar{C_t} = f_t * C_{t-1} + i_t * \bar{C_t} \tag{7}$$

The network-training is guided by CTC loss function. At each time step, the CTC is used to back-propagate for predicting the error vector in the duration of training. Only the corresponding ground-truth (GT) text and the output matrix of the NN are fed to the CTC lost function. CTC's output is a probability distribution over all label sequences on the given input sequence. For minimizing an objective function inspired by HMM backward-forward algorithm, the loss function tries all possible alignments of the GT text including in the image. Then, it takes the totality of all scores. Thus, the score of a GT text is high when the totality of the alignment-scores has a high value. The conditional probability formula is briefly described as shown in equation (8):

$$p(\pi|x) = \prod_{t=a}^{T} y_{\pi_t}^t, \forall\pi \tag{8}$$

In equation (8), x is the input of text sequence, π represents a possible output of text sequence and $y_{\pi_t}^t$ is the probability (a given label) from π at time t.

A function F is used to operate the match sequences from previous step π for producing the final output text sequences l by deleting blank labels and combining repeated labels. The example of the operation of the function F over a possible sequence is shown in equation (9).

$$F\ (aa\_\_c\_bb) = (acb) \tag{9}$$

There may be more sequence π that corresponds to a single final sequence. Thus, the probability (for a final sequence) l is calculated by using equation (10).

$$p(l|x) = \sum_{\beta(\pi)=l} p(\pi|x) \tag{10}$$

For each input text x and sequence l, the maximum probability $p(l|x)$ is the final output.

IV. DESIGN OF PROCESS FLOW

The design of process flow for this recognition system is shown in Fig. 3. Since this system is to recognize the student ID card, one of the cards is inserted into it as input.

The operation of this recognition system has four steps:

- Step 1: The text regions including in the card are localized by using Faster R-CNN.
- Step 2: The convolution features are extracted from the recognized text regions.
- Step 3: The extracted features are arranged according to the timestamps.

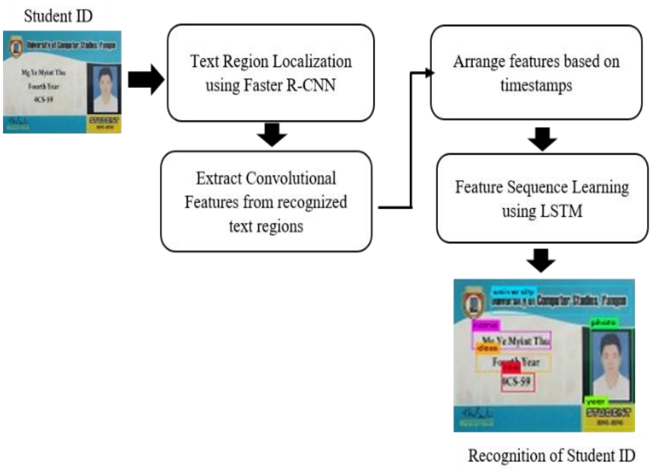

Fig. 3. Design of process flow

- Step 4: The arranged features are learned sequentially by using LSTM.

After passing through these four steps, a recognized card is obtained as the output of this system.

The processing of text region localization and sequence learning are also described in detail in the following sub-sections.

A. *Text Region Proposal Generation*

In this paper, the text regions of the target six classes are identified for training in the regional proposal network (RPN). The target names of these six classes are "**Name**", "**Roll Number**", "**Class**", "**Photo**", "**Year**", and "**University**".

To recognize text regions of target classes, the annotation task is processed over student ID cards before training. For training, the labelling tool is used to annotate the regions of the target classes. Fig. 4 shows the process of region annotation to train in the regional proposal network. The labelling tool produces a file as its output. The file types of the output are "txt" or "xml". The contents including in the output file are formatted with the fields such as "<target class id>, <x>, <y>, <width>, <height>". The sample output file is shown in Fig. 5.

After annotation from the dataset of the ID cards, the training process is ready for text region localization. For training, 100 student ID cards are collected.

Fig. 4. Process of annotation

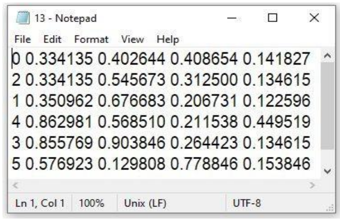

Fig. 5. Annotation results

Then, the collected card dataset is augmented into 300 ID cards by using Data Augmentation techniques. These techniques support the practitioners by increasing the diversity of data that can be used for training models without the actual new data collection.

Since the commonly used techniques in large neural networks training are cropping, padding, and horizontal flipping, this system used the combination of these techniques for data augmentation to increase the training data. Moreover, this system also used colour and grey scale image for training. The criteria used in data augmentation are shown in Table I.

The result of Regional Proposal Network after training is shown in Fig. 6.

B. *Text Sequence Learning and Recognition*

LSTM model is used to learn text sequences from the recognized text regions. Firstly, 70000 printed text images (including words and phrases) are collected and resized into 32 pixels height and 128 pixels width for training. Then, text images are split into 75% training dataset and 25% test dataset.

Before learning text sequences, features are extracted from the recognized text regions by using Convolutional Neural Network (CNN). Three Convolution layers are used in this system and text image features that are extracted layer by layer from the training data to learn feature sequences. The structure of these Convolution layers is shown in Table II.

TABLE I. DATA AUGMENTATION CRITERIA

Criteria		Value
Rotation	Max left degree	10
	Max right degree	10
Random Noise	Probability	0.5
Blur	Probability	0.1
Vertical flip	Probability	0.2
Horizontal flip	Probability	0.2
Resize		416 x 416

Fig. 6. Results of RPN

TABLE II. CONVOLUTION LAYERS OF FEATURE EXTRACTION

Layer (type)	Output Shape	Param #
the_input (InputLayer)	(None, 32, 128, 1)	0
conv2d_3 (Conv2D)	(None, 32, 128, 128)	1280
activation_3 (Activation)	(None, 32, 128, 128)	0
max3 (MaxPooling2D)	(None, 16, 64, 128)	0
conv2d_4	(None, 16, 64, 128)	147584
activation_4 (Activation)	(None, 16, 64, 128)	0
max4 (MaxPooling2D)	(None, 8, 32, 128)	0
conv2d_5 (Conv2D)	(None, 8, 32, 256)	0
Total params		444,032
Trainable params		444,032
Non-trainable params		0

The visualization of feature maps from layer to layer is shown in Fig. 7. After extracting features maps from convolutional neural network, the next step is to learn feature sequences for LSTM training. The feature maps are reshaped into 32 x 256 and then passed these features through LSTM unit that is bidirectional. For the shape feature with 32 x 256, each of the 256 features in 32-time steps is inserted as input for the respective time step.

The output probability values for ground truth characters are produced after passing through the final dense layer. Then, the softmax activation is used to extract the corresponding output character at each of the 32-time steps. After performing softmax activation, the process of decoding Connectionist Temporal Classification (CTC) is performed. CTC is used to address the problem of not knowing the mapping of different parts of images to different characters.

For each input sequence l , and text x, the final output is the maximum probability $p(l|x)$. This best path CTC decoding is shown in Fig. 8. The 32-time steps of decoding activation maps and its corresponding predictions are shown in Fig. 9. To align the recognized text sequences, CTC's best path encoding is used to select the best path from timestamp and the most probable ground truth label indexes are selected from last activation layer of deep network by passing through each timestamp and then convert those indexes into character labels as the recognized outputs. The number of ground truth character indexes is depending on distinct characters from the collection of text images.

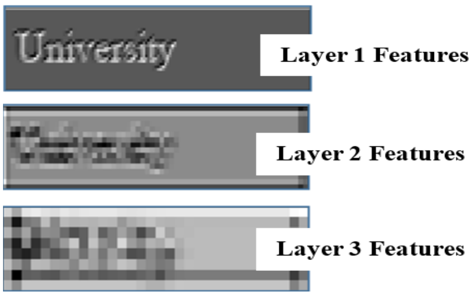

Fig. 7. Layer to Layer Feature maps extracted from CNN

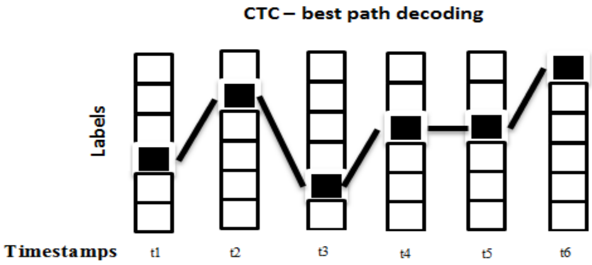

Fig. 8. Best path CTC decoding

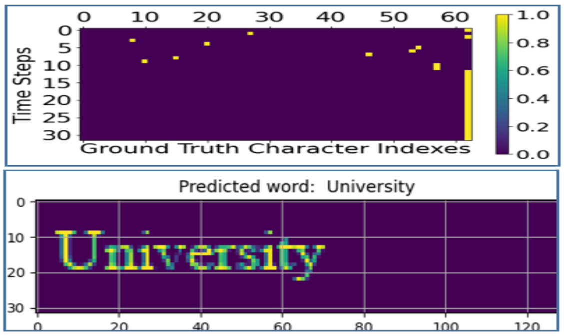

Fig. 9. Activation map for 32 steps decoding and its predicted output

V. EXPERIMENTS

The experiments are conducted in two processes for training and testing – Regional Proposal Network for text region localization and LSTM for sequence learning and recognition. Both processes are independent for training before real time testing. Using the deep learning approaches, two types of training are experimented for testing the performance accuracy of real-time text recognition.

One of the type of experiments for training is to determine where text regions are in ID card images. Moreover, another type of training is also used to learn text sequences by feature extraction from Convolutional Neural Network.

A. *Experimental Setup*

For conducting the training experiments, the experimental environment is setup by using MSI gaming laptop with Intel(R) Core (TM) i7-10875H CPU @ 2.30GHz, a NVIDIA GeForce RTX 2070 8 GB with Max-Q Design GPU and 16 GB memory.

B. *Experimental Results*

Firstly, training data are collected for text region localizations with Regional Proposal Network. Since RPN model has two outputs – classifier and regressor, the two loss functions are applied to both the RPN model and Classifier model. The losses of regressor are localization loss (Localization Loss or Bounding Box regressor Loss for the RPN) and objectness_loss (Loss of the Classifier that classifies if a bounding box is an object of background or interest). The losses for the Final Classifier are classification loss (Loss for the classification of detected objects into various classes) and localization loss (the Bounding Box regressor Loss or Localization Loss).

The results of losses obtained from training the RPN network by using an augmented 300 ID dataset with the maximum iteration of 2k. During the training of regional proposal network, the values of loss classes (Objectness, Localization, Classification) are gradually approached to zero by tuning the number of epochs. By decreasing RPN's losses, a number of bounding boxes called Region of Interests (ROIs) that has high probability of containing text regions are generated. For testing ID cards for test region localization, new 20 student's ID cards are used. The accuracy of the testing is 95%. For getting higher accuracy, more datasets and iterations are needed in training.

The experimental results of text sequence learning and recognition are produced by collecting 70000 printed text images and splitting them into two parts: 75% for training and 25% for testing. Before text sequence learning with LSTM model, the convolutional features are extracted from text images. Moreover, two types of procedures are conducted for feature extraction before learning text sequences to produce their comparison results. One approach is to pass the whole text image into CNN model and extract features once for the whole image. Another approach is sliding text images into horizontally and extracting features for each slide regions of text image. Finally, feature sequence of all slice regions to input are combined into LSTM sequence model. The recognition accuracies from learning these two features over LSTM model are shown in Table III.

According to the experimental analysis of sequence learning with LSTM model, the feature extraction methods are also impacted on learning sequence model. The font scale variant of characters in learning sequence model is one of the weakness in this model. To overcome this weakness, large dataset with different scale font is needed for deep learning training. Moreover, it can be overcome by making the contribution of font invariant features for different scale of character when creating the application in OCR model.

VI. CONCLUSION

Recently, computer vision and image analysis fields have been increasingly used deep learning because it provides better accuracy than many traditional approaches. In this paper, an approach to identify Student ID cards by using deep learning (Regional Proposal Network and LSTM) is presented. The aims of this approach are to overcome the difficulty in extracting text region from the ID cards by using Faster R-CNN because some documents are already predefined the formats and layouts of the text. By using the regional proposal network, text regions areas can be easily identified. Moreover, for text recognition, sequence learning model (LSTM) are used to avoid segmentation problems of OCR application. This paper described English Text recognition. The recognition of Myanmar Text (printed and handwritten) will be the further extension of this paper.

TABLE III. COMPARISON OF TEXT RECOGNITION ACCURACIES FROM LSTM MODEL

Total Number of Printed Text Image (70000)	#Training	#Testing	Training accuracy (%)	Testing accuracy (%)
Convolution Feature Type I (For the whole image)	52500	17500	92	90
Convolution Feature Type II (with sliding window)	52500	17500	95	93

REFERENCES

[1] T M. Breuel, A. Ul-Hasan, M A. Al-Azawi, F. Shafait, "High-performance OCR for printed English and Fraktur using LSTM networks", In *2013 12th international conference on document analysis and recognition*, 2013 Aug 25, pp. 683-687, IEEE.

[2] A S. Graves, "Offline handwriting recognition with multidimensional recurrent neural networks", *Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada*, 2009, pp. 545:552.

[3] G A. Abandah, F T. Jamour, E A. Qaralleh, "Recognizing handwritten Arabic words using grapheme segmentation and recurrent neural networks", *International Journal on Document Analysis and Recognition (IJDAR)*, 2014 Sep 1,17(3), pp. 275-91.

[4] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition", *IEEE transactions on pattern analysis and machine intelligence*, 2008 May 30, 31(5), pp. 855-68.

[5] B. Shi, X. Bai, 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*, 2016 Dec 29, 39(11), pp. 2298-304.

[6] H. Zhang, H. Wei, F. Bao, G. Gao, "Segmentation-free printed traditional Mongolian OCR using sequence to sequence with attention model", In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017 Nov 9, Vol. 1, pp. 585-590, IEEE.

[7] S. Yousfi, S A. Berrani, C. Garcia, "Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos", In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015 Aug 23, pp. 1026-1030, IEEE.

[8] C. Zhai, Z. Chen, J. Li, B. Xu, "Chinese image text recognition with BLSTM-CTC: a segmentation-free method", In *Chinese Conference on Pattern Recognition*, 2016 Nov 5, pp. 525-536, Springer, Singapore.

[9] J. Zhu, H. Ma, J. Feng, L. Dai, "ID card number detection algorithm based on convolutional neural network", In *AIP Conference Proceedings*, 2018 Apr 18, Vol. 1955, No. 1, pp. 040124, AIP Publishing LLC.

[10] A F. Mollah, S. Basu, N. Das, R. Sarkar, M. Nasipuri, M. Kundu, "Text region extraction from business card images for mobile devices", arXiv preprint arXiv1003.0642, 2010 Mar 2.

[11] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", arXiv preprint arXiv:1506.01497, 2015 Jun 4.

[12] Faster R-CNN [Online]. Available: https://medium.com/@smallfishbigsea/faster-r-cnn-explained-864d4fb7e3f8.

[13] Faster R-CNN for Object Detection [Online]. Available: https://towardsdatascience.com/faster-r-cnn-for-object-detection-a-technical-summary-474c5b857b46.