Computers, Materials & Continua
DOI:10.32604/cmc.2021.018724
Article

Tech Science Press

Recurrent Convolutional Neural Network MSER-Based Approach for Payable Document Processing

Suliman Aladhadh[1], Hidayat Ur Rehman[2], Ali Mustafa Qamar[3,4,*] and Rehan Ullah Khan[1]

[1]Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia
[2]Ainfinity Algorythma, Abu Dhabi, United Arab Emirates
[3]Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
[4]Department of Computing, School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
[*]Corresponding Author: Ali Mustafa Qamar. Email: al.khan@qu.edu.sa

Received: 19 March 2021; Accepted: 20 April 2021

Abstract: A tremendous amount of vendor invoices is generated in the corporate sector. To automate the manual data entry in payable documents, highly accurate Optical Character Recognition (OCR) is required. This paper proposes an end-to-end OCR system that does both localization and recognition and serves as a single unit to automate payable document processing such as cheques and cash disbursement. For text localization, the maximally stable extremal region is used, which extracts a word or digit chunk from an invoice. This chunk is later passed to the deep learning model, which performs text recognition. The deep learning model utilizes both convolution neural networks and long short-term memory (LSTM). The convolution layer is used for extracting features, which are fed to the LSTM. The model integrates feature extraction, modeling sequence, and transcription into a unified network. It handles the sequences of unconstrained lengths, independent of the character segmentation or horizontal scale normalization. Furthermore, it applies to both the lexicon-free and lexicon-based text recognition, and finally, it produces a comparatively smaller model, which can be implemented in practical applications. The overall superior performance in the experimental evaluation demonstrates the usefulness of the proposed model. The model is thus generic and can be used for other similar recognition scenarios.

Keywords: Character recognition; text spotting; long short-term memory; recurrent convolutional neural networks

1 Introduction

Deep Learning (DL) relies on the powerful function approximation and representation attributes of deep neural networks [1]. DL's innovation and realization have revolutionized many areas, including computer vision, speech recognition, pattern recognition, and natural language processing. DL has enabled computational mathematical models and frameworks, which comprise multiple interlinked processing intermediate layers, to learn the inherent representations of data.

This learning is achieved with multiple levels of abstraction by introducing multiple layers [2]. Recognition of sequence objects, such as handwritten text, scene text, and the musical score, is challenging compared to other similar problems. The challenge comes from the prediction of the series of object labels rather than single labels. The second challenge to sequence-based objects is their arbitrary lengths. The lengths of sequence objects may vary on a case-to-case basis, and no restrictions can be imposed as these occur in natural problems and represent natural circumstances. An example of such sequence objects in the scene is the text with the word "yes" with only three characters and a word like "investments" having eleven characters. This poses a challenge to detection and recognition algorithms. State-of-the-art contains efforts to address this problem using conventional and non-conventional Machine Learning (ML) approaches. DL is used in object detection, image understanding, document analysis, and text recognition.

In this paper, we propose an end-to-end Optical Character Recognition (OCR) system, which does both localization and recognition and serves as a single unit to automate payable document processing such as cheques, vendor invoices, and cash disbursement. For text localization, the maximally stable extremal region is used, which extracts a word or digit chunk from an invoice. This chunk is later passed to the deep learning model, which performs text recognition. The deep learning model utilizes both the convolution neural network and long short-term memory (LSTM). The convolution layer is used for extracting features, which are fed to the LSTM. The model integrates feature extraction, modeling sequence, and transcription into a unified network. The proposed architecture, being an end-to-end trainable network, handles the sequences of unconstrained lengths, independent of the character segmentation or horizontal scale normalization.

Furthermore, it applies to both the lexicon-free and lexicon-based text recognition, and finally, it produces a comparatively smaller model, which can be implemented for practical applications. The overall superior performance in the experimental evaluation demonstrates the usefulness of the proposed model. The model is thus generic and can be used for other similar recognition scenarios.

The rest of the paper is organized as follows. Section 2 presents the related work, whereas the proposed architecture is given in Section 3. Similarly, Section 4 discusses the experimental analysis and evaluation, whereas Section 5 concludes the paper.

2 Related Work

Shi et al. [3] investigated text recognition in the scene. A unified framework and novel deep architecture are presented that integrate feature extraction, sequence modeling, and transcription. The proposed approach is end-to-end trainable and handles sequences without restrictions on the length. The approach is also independent of the prior lexicon and generates a comparatively smaller model for real-time, real-world scenarios. Tian et al. [4] propose an approach for text localization in natural images. They term the approach as the Connectionist Text Proposal Network (CTPN). CTPN is based on the vertical anchor that efficiently predicts text location and the scoring of text and non-text for fixed-width proposals. The approach fuses the Recurrent Neural Network (RNN) with the Convolutional Neural Network (CNN). The CTPN, which uses RNN and CNN, is shown to work reliably on multi-scale and multi-language text. The approach does not need the post-processing steps compared to the previous approaches.

In [5], the authors propose an approach to text detection in complex scenarios involving panorama images. The approach exploits the Extremal Regions (ER) as well as the fusion of edge information, probabilistic color detection, and geometric properties for segmenting the text from the background. The authors report good overall detection performance. In [6], the authors present a novel approach for detecting tables in document images. The workflow for table detection is based on three unique steps: the first one is preprocessing, followed by the detection of horizontal and vertical lines. The last one is the table detection based on the previous two steps. The performance is evaluated using forms, magazines, newspapers, scientific journals, certificates, handwritten documents, and cheques. In [7], the authors use morphological operators for text feature extraction for text line segmentation in documents. The algorithm is based on projecting multiple histograms. From the horizontal projection on the text image, line segments are extracted based on the peak horizontal projection. Threshold-based segmenting segments the images into multiple parts. The histogram's vertical projection is exploited for the line segments, followed by the decomposition in words and, finally, characters using different thresholds. They got an accuracy of 98%.

Yang et al. [8] propose a hierarchical network based on two unique characteristics. First, the proposed network follows the structure of the documents. Second, the approach employs two attention levels that are applied at the sentence and word levels. In evaluating six large-scale text tasks, the proposed method outperforms the state-of-the-art by a large margin. The approach in [9] investigates Neural Network (NN) architecture for multi-label text classification tasks. The article proposes that simple NN models and the integration of rectified, dropout, and AdaGrad are suitable for this task. Specifically, Backpropagation for Multi-Label Learning's (BP-MLL) ranking loss minimization is useful to be replaced with the commonly used Cross Entropy Error (CEE) function. The evaluation suggests that rectified linear units (RLU), dropout, and AdaGrad outperform other detection approaches based on six large-scale text datasets. Graves et al. [10] proposed an approach based on the RNN that is specifically designed for sequence labeling tasks, in which the data is complex and contains multi-range, directional dependencies. The proposed network is robust to the size q of the lexicon. The impact of hidden layers and the use of hidden layers' context is also demonstrated. The approach significantly outperforms the state-of-the-art.

In [11], two text extraction approaches from the natural images are compared based on the edge-based and connected-component. Furthermore, DL and RNN are also widely used for generic object detection in images. In [12], Zuo argues that CNN networks alone are not adequate and suitable to model the complex relationship between pixels in images. RNN, on the other hand, can model the inherent contextual dependencies in digital images. Therefore, the authors propose to merge CNN and RNN, especially for tasks involving pixel-based detection. As an example application, the work demonstrates its use of the fusion approach for skin color detection in two datasets. The work in [13] is also based on the similar concept of using CNN and RNN together due to the inherent complexities involved in the images' objects. The proposed approach is termed the CNN-RNN framework. It uses image-label embedding to learn the semantic label interdependency and the relevance of the image label. The CNN-RNN is end-to-end trainable. The approach outperforms the state-of-the-art.

In [14], the authors also propose a recurrent CNN represented as the RCN for image-based object recognition tasks. The activities of the proposed Recurrent Convolutional Neural Network (RCNN) layers and units evolve by modulated activities of the neighboring units, thus learning the contextual information. On evaluating the proposed approach using the four datasets, the RCNN outperforms state-of-the-art models on all of these datasets and demonstrates the advantage of RCNN. Elfwing et al. [15] propose the deep architecture of the Free-Energy Restricted Boltzmann Machines (FE-RBM). The RBMs are stacked on top of each other, and the class node is connected to all the hidden layers to improve performance. The performance of the approach shows its effectiveness.

3 Proposed Methodology

OCR generally requires two steps: the first is the localization, and the second being recognition. Our method of processing payable documents is also divided into two steps. The first step is text localization. In this step, an image is segmented to get only the text candidate region, and other regions are removed. These segmented regions or chunks are then passed to the text recognition module, which then transcribes the image. We use the Maximally Stable Extremal Region (MSER) for text localization due to its robustness to noise and illumination. MSER detects text chunks, which are then passed on to the recurrent convolutional neural network, which generates a transcription of the image.

3.1 Maximally Stable Extremal Region (MSER)

MSER is proposed by Matas et al. [16] for the correspondence between two image-based objects with different viewpoints. The MSER regions are used for blob detection in digital images. MSER regions possess two properties: first, they are affine invariant and are independent of warping or skewness. Secondly, the regions are sensitive to lightness or darkness. The intensity function calculates the MSER regions in the corresponding region and the outer boundary, which results in the regions' valuable characteristics for detection tasks. Sambyal et al. [17] proposed character segmentation and text extraction based on the MSER. The MSER is used to treat the essential letter candidates. The MSER threshold regions are used to determine the various connected components for various characters' identification. The algorithm is evaluated on the character sets from English, Russian, Urdu, and Hindi languages. The authors report good performance for the English and Russian languages characters, but comparatively low performance for the character set of Urdu and the English languages. The authors advocate the simplicity and less overhead of the proposed approach. In [18], Sung et al. propose the Extremal Region (ER) tree construction. It is advocated that the use of MSER regions alone, as done by Sambyal et al. [17], is not a viable solution due to the strict requirements of maximum stability and, therefore, achieves decreased performance. The approach employs sub-path sampling, pruning, character candidate selection, and finally, using Adaptive boosting to verify the candidates in extracted characters. Thus, the approach achieves an increased recall of 8%, precision of 1%, and F-measure of 4%. In [19], the authors propose a multi-level MSER for text segmentation from the scenes. The proposed approach defines a segmentation score based on the four measures of stroke width, boundary curvature, character confidence, and color constancy. The best MSER scored from each channel is fused for final segmentation. In [20], the authors propose a text detection approach based on enhanced MSER. The approach employs an enhancement based on edge detection and is termed as edge-enhanced MSER for basic letter candidates. Based on the geometric and stroke width information, the basic letters' candidates are then filtered for excluding non-textual objects. Finally, the letters are paired and subsequently separated into discrete words using the text lines' identification.

In our paper, an image is fed to MSER, which extracts the character candidate region and the input image's noise. Further enhancement to the MSER is done to extract only the characters and discard the non-text regions such as logos, lines, and boxes. The text and text regions are separated based on the stroke width of the candidate region. The characters usually have a stroke width less than that of non-text. After getting the character candidate region, chunking of individual letters was done to form words from characters. Word chunks are formed if two letters overlap each other in a horizontal direction.

3.2 Proposed MSER Recurrent Convolutional Neutral Network (MRCNN)

The proposed architecture is shown in Fig. 1. Our network consists of four units, the MSER layer, convolution layers, recurrent layers, and the transcription layers. The MSER is applied as a preprocessing to segment the characters. The convolutional layers extract and learn the feature sequence from each input image character. The recurrent network is constructed from the feature sequence of the convolutional layer output, making a probabilistic prediction for each frame. The final layer of transcription that gets input from the recurrent layer is used to translate the recurrent layer predictions into sequence labels. The convolutional and the recurrent networks are jointly trained with one loss function.

3.3 Convolutional Feature Map

In the proposed approach, the convolutional layers are like conventional Deep CNN, using the convolutional and max-pooling layers and removing the fully connected layers. This setup thus extracts sequential features from input. One constraint is the similar scaling of the input images. The feature vectors of feature maps are produced by applying the convolutional layers used by the next recurrent layer. Each feature vector is generated in a left to right fashion on the feature map using a column. The width of the column is kept fixed, i.e., 26 pixels. In the proposed network, deep features are conveyed into sequential representations invariant to the length of sequence-like objects. The translation invariance comes from applying the layers of convolution, max pooling, and element-wise activation function operating on the local regions. Thus, the feature maps' columns represent the original image's rectangle region, also referred to as the receptive field.

3.4 LSTM Recurrent Labeling

For predicting the label Y for each frame x in feature sequence X, a deep bi-directional RNN is constructed on top of the CNN layers as the recurrent layers. The RNN has strong capabilities in capturing the context information in the image sequence. It can trace back the errors to input convolutional layers that calculated these features, allowing a joint CNN and RNN to train in sequence. Furthermore, RNN can operate on arbitrary length sequences. A basic unit of the RNN contains a self-connected hidden layer in between the input layer and the output layer. When it receives a frame in the input sequence, it issues an update to its internal state using a non-linear function. This function takes the current input and previous state as inputs and predicts the current class. The generic RNN networked units suffer from the vanishing gradient problem, as discussed by Bengio et al. [21]. Thus, this problem adds a burden on the overall training setup and reduces the range of context storage. Thus, the LSTM of [22,23] as RNN types addresses this problem. Since the context from both directions is valuable and complementary to each other, we follow [3,24], combining two LSTMs. We combine a forward and backward LSTM for two-way directional LSTM.

Moreover, multiple LSTMs layers can be used to construct a deep LSTM model. The deep LSTM has contributed to the task of speech recognition [24]. Furthermore, it allows for a higher level of abstractions than simple LSTM. The error is propagated in the opposite directions of the directional LSTM using the Back-Propagation Through Time (BPTT). In the last stages of RNN, the propagated sequence differentials are mapped. This inverts the operation of feature maps conversion into feature sequences, fed to the convolutional layers.

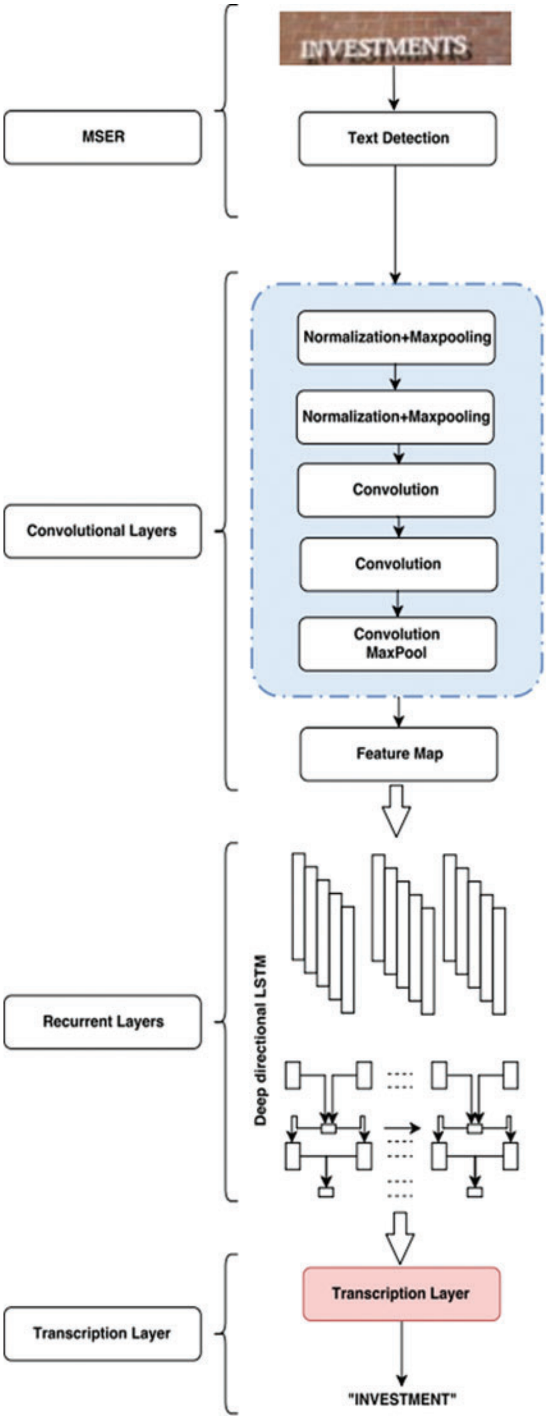


Figure 1: The proposed architecture of the CNN-LSTM network. The network contains three layers: the convolution layer, which learns and thus extracts features, LSTM Recurrent layer that predicts the class label for each frame; and the transcription layer that maps the predictions into the final label

3.5 Label Transcription

The frame-based predictions of the RNN are converted into sequence labels. This process is termed transcription, where we find the label sequence based on the highest probabilistic distribution of predictions. There are two kinds of transcriptions: lexicon-based and lexicon-free transcriptions. A lexicon puts a constraint on the predictions. In a lexicon-free setup, the predictions are unconstrained. In lexicon mode, the highest probability drives the predictions for label sequence. The transcription is done using the connectionist temporal function, which uses a forward-backward algorithm for finding the optimal candidate. That is why it learns the contextual information about predicting a chunk. However, the Back-Propagation Through Time (BPTT) is applied in the recurrent layers for calculating the error differentials in these layers.

4 Experiments and Results

This section contains the details of the experiments along with a detailed discussion of the obtained results.

4.1 Network Training

We consider an image training dataset $X = \{I, L\}$, where I represent the images, and L stands for the ground truth labels. The network minimizes the negative log-likelihood of the conditional probability of the ground truth. A Stochastic Gradient Descent (SGD) is used and calculated by the back-propagation algorithm for training the network. The "forward-backward" algorithm of [25] propagates the error differences backward in the Transcription layer. For per-dimension learning rates calculation and its optimization, we used the ADADELTA algorithm of [26]. Compared to others, the ADADELTA automatically calculates the learning rates. We also found that the optimization by the ADADELTA converged faster.

4.2 Network Configuration

The architecture of the convolution setup is extended from the work of Simonyan et al. [27]. Tab. 1 shows the configuration of the network. The network of [27] is adapted to work for the English text. As such, in the max-pooling layers of row 8 and row 11 (Tab. 1), we adapted pooling strides of 1 × 2 compared to the conventional 2 × 2. We also represent them as third and fourth max-pooling layers. This results in the feature maps having larger widths, thus producing a comparatively more extended feature sequence. Our network has deep convolutional layers with deep recurrent layers and uses the batch normalization technique introduced by Ioffe et al. [28]. The technique in [28] is beneficial for training a network of these extreme depths. The network is augmented with the batch normalization layers. These layers are inserted after the third, fifth, and seventh convolutional layers. The batch normalization process of [28] greatly reduces the training times, thereby expediting the network's execution.

4.3 Environmental Specification and Experimental Details

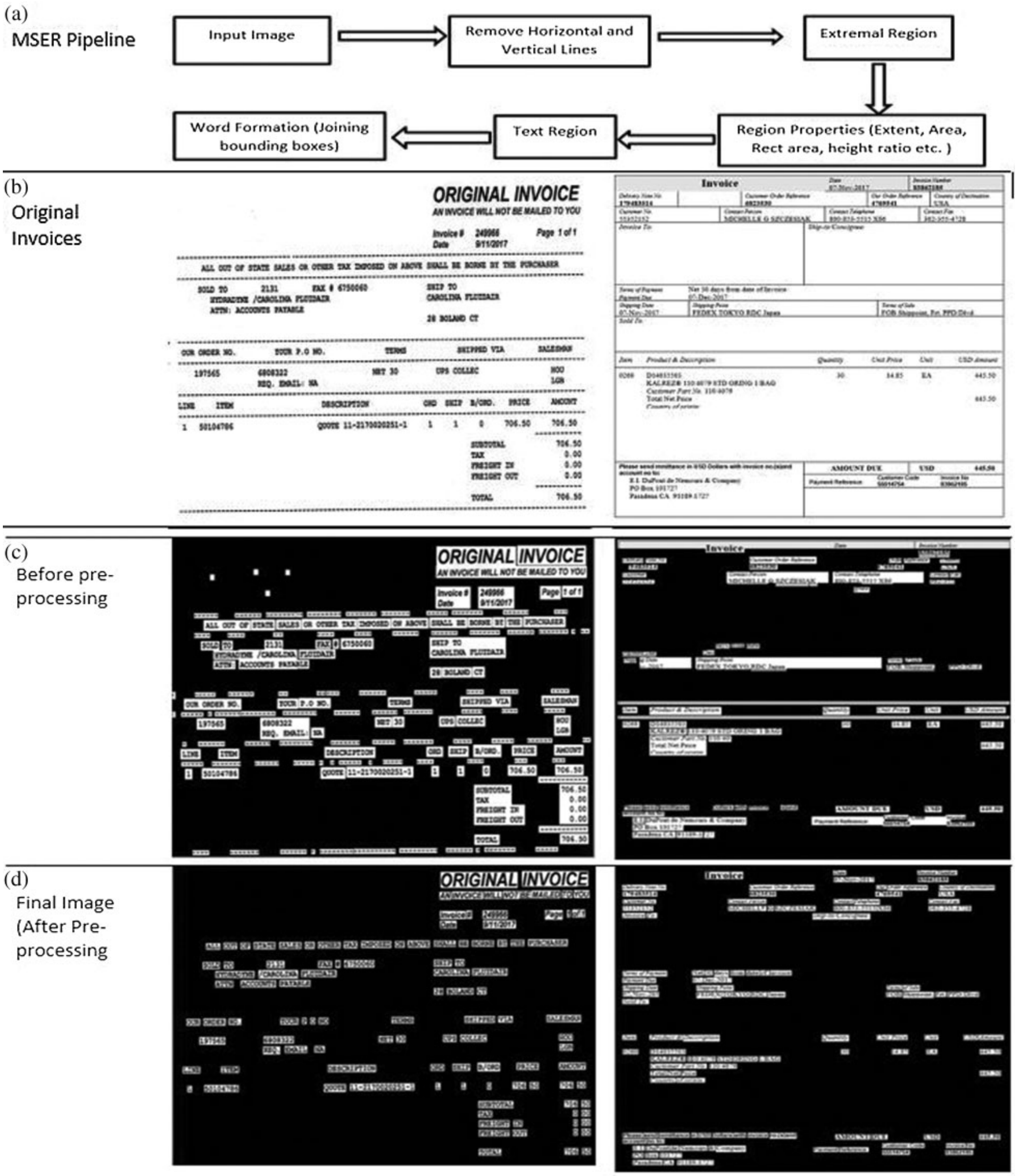
We used the GeForce GTX 1080 Ti server containing 3584 cores and 12 GB of dedicated GPU memory for the experimental setup. Besides the GPU configuration, the server contains 20 CPUs and 32 GB of memory. The model was trained on 0.65 million images and validated against 0.15 million images. Due to the simplicity of the training data (since the data only consists of numbers and alphabets), the model converged in just two hours. On a test set of 20 thousand payable cash vocabulary, the model reported an accuracy of 95 percent when used without lexicon. When used with a lexicon, the model reported 99 percent accuracy in predicting words. One of the advantages of this network is the duality of its configuration, i.e., it can be used to predict words restricted either to a particular lexicon or without any restrictions. Restricting the model prediction to a lexicon helps in the correct identification of key terms that are used in various financial and analytic processes. In this way, we utilize both lexicon-based and lexicon-free models. Predicting words through the lexicon-based model is cost-effective, but if there is a small lexicon, e.g., twenty thousand words and CUDA environment, then the time taken by the lexicon-based model does not cost that much.

Table 1: RNN configuration

Layer	Details
Input data	100 × 32 image
Convolution	(1:64, 3 × 3, 1, 1, 1, 1)
Max pooling (1)	(2 × 2, 2, 2)
Convolution	(64:128, 3 × 3, 1, 1, 1, 1)
Max pooling (2)	(2 × 2, 2, 2)
Convolution	(128:256, 3 × 3, 1, 1, 1, 1)
Convolution	(256:256, 3 × 3, 1, 1, 1, 1)
Max pooling (3)	(2 × 2, 1, 2, 1, 0)
Convolution	(256:512, 3 × 3, 1, 1, 1, 1)
Convolution	(512:512, 3 × 3, 1, 1, 1, 1)
Max pooling (4)	(2 × 2, 1, 2, 1, 0)
Convolution	(512:512, 2 × 2)
Transpose-split-sequential	
LSTM-bidirectional	Hidden: 256 units
LSTM-bidirectional	Hidden: 256 units
Transcription	

4.4 Experimental Results and Discussion

Our deep pipeline consists of two phases: the first phase is image detection, i.e., detection of textual geometries in a scanned document, and the second one is the transcription of detected textual geometry. MSER, already discussed in this paper's methodology section, is widely used for text detection [17–20]. Although MSER detects the textual geometries accurately in scanned documents, it detects some non-textual regions such as horizontal and vertical lines, logos, noise (bar codes and dashes). MSER works in the following manner: firstly, it performs thresholding based on the luminance, and then it extracts the connected components called extremal regions that survive the recursive thresholding. Thus, the final regions we obtain are the maximally stable extremal ones. To get only the stable textual regions, we enhanced the MSER module to get only textual regions. Before passing the image to the MSER module, we performed some preprocessing to remove vertical and horizontal lines, as shown in Fig. 2.


(a)
MSER Pipeline
Input Image
Remove Horizontal and Vertical Lines
Extremal Region
Region Properties (Extent, Area, Rect area, height ratio etc.)
Text Region
Word Formation (Joining bounding boxes)
(b)
Original Invoices
(c)
Before pre-processing
(d)
Final Image (After Pre-processing

ORIGINAL INVOICE
AN INVOICE WILL NOT BE MAILED TO YOU

Invoice # 249966 Page 1 of 1
Date 9/11/2017

ALL OUT OF STATE SALES OR OTHER TAX IMPOSED ON ABOVE SHALL BE BORNE BY THE PURCHASER

SOLD TO 2131 FAX # 6750060
HYDRADYNE /CAROLINA FLUIDAIR
ATTN: ACCOUNTS PAYABLE

SHIP TO
CAROLINA FLUIDAIR

28 BOLAND CT

OUR ORDER NO.	YOUR P.O NO.	TERMS	SHIPPED VIA	SALESMAN
197565	6808322 REQ. EMAIL: NA	NET 30	UPS COLLEC	HOU LGN

LINE	ITEM	DESCRIPTION	ORD	SHIP	B/ORD.	PRICE	AMOUNT
1	50104786	QUOTE 11-2170020251-1	1	1	0	706.50	706.50

SUBTOTAL	706.50
TAX	0.00
FREIGHT IN	0.00
FREIGHT OUT	0.00
TOTAL	706.50

Figure 2: (a) Shows the overall MSER pipeline (b) shows original images (c) shows binary images after applying MSER (d) shows the final image after applying preprocessing and MSER

Lines are not the only type of noise that we encounter in documents, and there are some other types of noise as well, such as logos and barcodes. To remove such noise, we got statistical properties of these regions obtained from the connected component analysis. For each of the connected components obtained from MSER, we obtain the following region properties as shown in Eqs. (1)–(5):

$$\text{Aspect\_ratio} = \frac{\text{region\_width}}{\text{region\_height}}, \tag{1}$$

$$\text{Solidity} = \frac{\text{Area}}{\text{region\_width} * \text{region\_height}}, \tag{2}$$

$$\text{Height\_ratio} = \frac{\text{image\_height}}{\text{region\_height}}, \tag{3}$$

$$\text{Rectangular\_area} = \text{region\_width} * \text{region\_height}, \tag{4}$$

$$\text{Extent} = \frac{\text{Area\_of\_the\_region}}{\text{Rectangular\_area}}. \tag{5}$$

Based on the region properties, we applied adaptive thresholding to determine whether a particular region is textual or not. Fig. 3 shows the graphical illustration of our discussion.

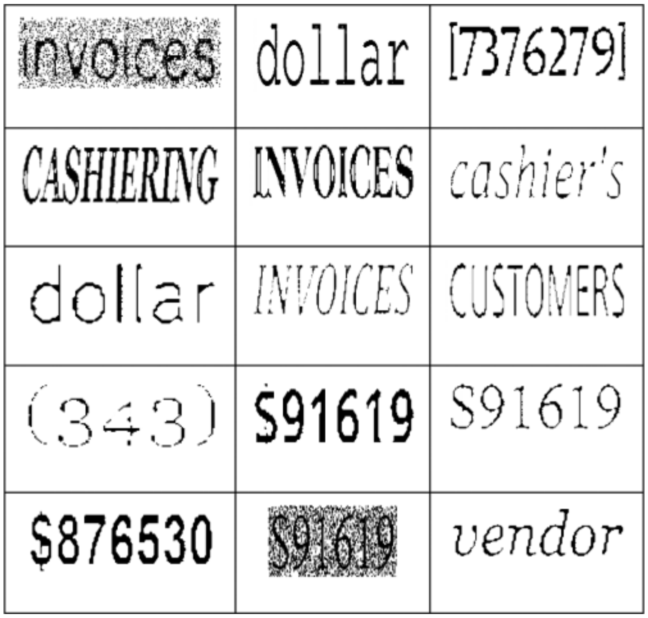


Figure 3: The description of the various transformations applied to the input data

After localizing the text candidate regions using MSER, we then pass the image to the RCNN (Recurrent convolutional neural network). The configuration of layers used for RCNN is discussed in the previous sections. RCNN uses the connectionist temporal classification, which predicts the whole word by using contextual information. However, it cannot help identify numbers since they do not rely on the context. We can use RCNN either in lexicon mode, i.e., predictions are constrained to a limited vocabulary, or lexicon-free mode, where the predictions are not constrained to any specific dictionary. We made changes in the RCNN and used the classifier in both the modes, i.e., lexicon mode and lexicon-free mode. The lexicon mode is used for predicting words that can later be used for extracting specific information from invoices, while the non-lexical mode is used for predicting alphanumeric and numbers. The connectionist cost function does not help predict numbers because number prediction does not depend on the previously predicted number.

Nevertheless, bi-directional LSTM can capture the geometrical information about a particular symbol and classify that number/alphanumeric. The previous results obtained using the connectionist temporal classification have been state-of-the-art [3]. In our case, we obtained an F-score of 0.99 during testing of 0.15 M images in lexicon-free mode.

5 Conclusion and Future Work

This paper presented a deep learning model based on Convolutional Recurrent Neural Network (CRNN) and Long Short-Term Memory (LSTM) to automate the tedious task of payable document processing. CNN helps in feature extraction. The extracted features are then passed to the LSTM. The RNN part is explored to describe the context of scene text images and predict sequence-like objects' structured outputs. The primary benefit of this approach is that it can handle both lexicon-free and lexicon-based text recognition. Furthermore, Maximally Stable Extremal Regions (MSER) were used for text extraction while avoiding noise. Our approach was able to get an accuracy of 95 percent on a test set of 20,000 payable images when used without lexicon. On the other hand, we got 99 percent accuracy while using the lexicon-based approach. In the future, we plan to use Bidirectional Gated Recurrent Units (BGRU).

Funding Statement: Researchers would like to thank the Deanship of Scientific Research, Qassim University, for funding publication of this project.

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

References

[1] K. Arulkumaran, M. P. Deisenroth, M. Brundage and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[2] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] 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, 2017.

[4] Z. Tian, W. Huang, T. He, P. He and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. European Conf. on Computer Vision, Lecture Notes in Computer Science*, Amsterdam, The Netherlands, vol. 9912, pp. 56–72, 2016.

[5] Y. Liu, K. Zhang, J. Yao, T. He, Y. Liu *et al.,* "An efficient method for text detection from indoor panorama images using extremal regions," in *Proc. IEEE Int. Conf. on Information and Automation*, Lijiang, China, pp. 781–786, 2015.

[6] B. Gatos, D. Danatsas, I. Pratikakis and S. J. Perantonis, "Automatic table detection in document images," in *Proc. Int. Conf. on Advances in Pattern Recognition*, Bath, United Kingdom, vol. 3686, pp. 609–618, 2005.

[7] N. Anupama, C. Rupa and E. S. Reddy, "Character segmentation for telugu image document using multiple histogram projections," *Global Journal of Computer Science and Technology*, vol. 13, no. 5, pp. 528–533, 2013.

[8] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola *et al.,* "Hierarchical attention networks for document classification," in *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, pp. 1480–1489, 2016.

[9] J. Nam, J. Kim, E. L. Mencía, I. Gurevych and J. Fürnkranz, "Large-scale multi-label text classification—Revisiting neural networks," in *Proc. Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, Springer, Berlin, Heidelberg, pp. 437–452, 2014.

[10] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke *et al.,* "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[11] S. Sharma, "Extraction of text regions in natural images," MS thesis, Rochester Institute of Technology, pp. 38, 2007.

[12] H. Zuo, H. Fan, E. Blasch and H. Ling, "Combining convolutional and recurrent neural networks for human skin detection," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 289–293, 2017.

[13] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang *et al.,* "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 2285–2294, 2016.

[14] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 3367–3375, 2015.

[15] S. Elfwing, E. Uchibe and K. Doya, "Expected energy-based restricted boltzmann machine for classification," *Neural Networks*, vol. 64, pp. 29–38, 2015.

[16] J. Matas, O. Chum, M. Urban and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[17] N. Sambyal and P. Abrol, "Automatic text extraction and character segmentation using maximally stable extremal regions," *International Journal of Modern Computer Science*, vol. 4, no. 3, pp. 136–141, 2016.

[18] M.-C. Sung, B. Jun, H. Cho and, D. Kim, "Scene text detection with robust character candidate extraction method," in *Proc. 13th Int. Conf. on Document Analysis and Recognition*, Tunis, Tunisia, pp. 426–430, 2015.

[19] S. Tian, S. Lu, B. Su and C. L. Tan, "Scene text segmentation with multi-level maximally stable extremal regions," in *Proc. Int. Conf. on Pattern Recognition*, Stockholm, Sweden, pp. 2703–2708, 2014.

[20] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk *et al.*, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Proc. IEEE Int. Conf. on Image Processing*, Brussels, Belgium, pp. 2609–2612, 2011.

[21] Y. Bengio, P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] F. A. Gers, N. N. Schraudolph and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, pp. 115–143, 2002.

[24] A. Graves, A.-R. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, pp. 6645–6649, 2013.

[25] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. on Machine Learning*, Pittsburgh, PA, USA, pp. 369–376, 2006.

[26] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," arXiv preprint arXiv: 1212.5701, 2012.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. on Learning Representations*, San Diego, CA, USA, pp. 1–14, 2014.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. on Machine Learning*, Lille, France, vol. 37, pp. 448–456, 2015.