Deep Neural Networks Combined with STN for Multi-Oriented Text Detection and Recognition

Saif Hassan Katper[1], Abdul Rehman Gilal[2]
Ahmad Waqas[4]
Department of Computer Science
Sukkur IBA University Sukkur, Pakistan

Aeshah Alsughayyir[5]
College of Computer
Science and Engineering
Taibah University, Madinah, KSA

Abdullah Alshanqiti[3]
Faculty of Computer and Information
Systems, Islamic University (IU), Madinah, KSA

Jafreezal Jaafar[6]
Center for Research in Data Science
Universiti Teknologi Petronas, Malaysia

Abstract—**Developing systems for interpreting visuals, such as images, videos is really challenging but important task to be developed and applied on benchmark datasets. This study solves the very challenge by using STN-OCR model consisting of deep neural networks (DNN) and Spatial Transformer Networks (STNs). The network architecture of this study consists of two stages: localization network and recognition network. In the localization network it finds and localizes text regions and generates sampling grid. Whereas, in the recognition network, text regions will be input and then this network learns to recognize text including low resolution, curved and multi-oriented text. Deep learning-based approaches require a lot of data for training effectively, therefore, this study has used two benchmark datasets, Street View House Numbers (SVHN) and International Conference on Document Analysis and Recognition (ICDAR) 2015 to evaluate the system. The STN-OCR model achieves better results than literature on these datasets.**

Keywords—Spatial Transformer Networks (STNs); Deep Neural Networks (DNN); ICDAR dataset; multi-oriented text; STN-OCR

I. Introduction

Text detection from scenary images is becoming focused area of research. It has attracted many researchers [1]–[3] from computer vision area due to its various applications such as tagging people in security cameras, understanding street signs for navigation, sign recognition in driver assisted systems, vehicle identifications, navigation people with low vision and processing bank cheques. In recent years, digital devices like smart phones or cameras are being used to produce a lot of multimedia contents (such as images and videos) across the world. It is now easy to capture the world's sceneries in the digital images through mobile devices as the prices are decreasing and performance is increasing. Not only the contents are generated at very large scale and easily uploadable but also accessed by billions of people on the internet. Many systems are developed to extract information for various purposes. However, the solutions are still under discussion with researchers.

Text detection and recognition from images in real world scenarios such as sign recognition in driver assisted systems, vehicle identification by reading license plates is major area of computer vision applications. Recent work [4], [5] presents Deep Neural Network for text detection with good results on horizontal text. However multi-oriented text is still lacking [6]. It is further discussed the very studies that text detection is a challenging task due to variations in text: orientation of text, text alignment, text visibility, multi-language text, low resolution or diversity of languages. Moreover, a recent research [7] is conducted for arbitrary text detection but still has some lackings in detecting multi-oriented text such as in object occlusion, large character spacing. These challenges remained continue even in state-of-art methods.

This study provides the solution for text detection and recognition problem from scenery images in arbitrary direction. Generally, there is no any restriction on text type in images, so if a human can read text whether of any type such as sign boards, calligraphy or newspaper then the systems should also detect and recognize text. The purpose of this work is all about giving generalized approach for multi-oriented text detection and recognition. Fig. 1 presents the abstract overview of the STN-OCR model.

Getting any information from scenery images is not simple task, it involves deep feature extraction. Many approaches [6]–[11] to this type of computer vision problems have been proposed. The research [12]–[15] in this area is mostly about end-to-end text recognition systems consisting two stages including text detection and text recognition. Text detection is reffered to finding text instances and highlighting textual part in images and text recognition means identifying that localized textual part of images. Text recognition stage evaluates that localized part of image and produces output in text form. Most of the existing work is focused on only one of these two stages either text detection or text recognition.

This paper is further divided into few sections for better presentation. For instance, the coming Section II discusses the related studies to ground the study need and value. Section III is about methods and methodologies. Section IV discusses the results and discussions obtained from the experiments. The last Section V concludes the study with certain remarks and recommendations.

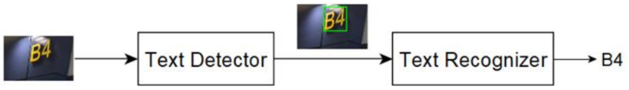


Fig. 1. Abstract Overview of STN-OCR Model.

II. Related Work

Text detection approaches focus on the first stage of two stages pipeline of scene text detection and recognition. It basically performs segmentation and produces words bounding boxes in scenery images.

Finding and localizing textual regions in complex backgrounds is really a challenging task and there are two approaches which overcome this task.

The first approach is character region which is used in Chen et al. [17], Epshtein et al. [18], Tian et al. [19], Yi et al. [20], Neuman et al. [12]–[14], [21]. Character regions based methods localizes character regions based on connected component analysis then characters are combined to form a word based on neighboring characters. Second approach is sliding windows which is used in Quack et al. [22], Anthimopoulos et al. [23], Posner et al. [24]. Sliding window based methods uses sliding window to find out textual regions within image and then uses machine learning techniques to recognize text.

A. *Traditional Approaches*

Traditional methods are mostly based on manual feature extraction in which human were involved to perform the scene text detection. These types of systems use features like stroke width transforms (SWT) [18], MSERs [12] and HOG-Features [7] to find textual regions and provide output to next stage which is text recognition. For text recognition, multiple approaches can be used to recognize textual regions like sliding window classifiers [25], ensembles of support vector machines (SVM) [26], KNN classifiers using HOG features [27]. Limitation in these methods is that these need expert knowledges to achieve best results.

The method proposed by Yuanwang et al. [16] is Exhaustive Segmentation (ES) for text detection. In their study, with the help of ES, character portions are extracted from image and filtering out non-character regions using two-layer filtering. These both are performed in parallel and support vector machines (SVM) classifier is used finally to cut out text regions. This method covers low resolution, blurred and small sized texts. ICDAR 2013 and Street View Text (SVT) datasets are provided for evaluating the performance of ES [16] approach. The ES method still has shown some lacking such as Broken Strokes, low resolution and dot-matrix fonts as shown in Fig. 2. In Aneeshan et al. [8], a novel approach has been proposed for multi-oriented text detection in images. In their study, Fourier Laplacian filtering is used for textual portions identification and then applied maximum-difference map separating image into text and non-text regions. In the end, Hidden Markov Model (HMM) is used for verification of selected text portions in image and non-textual regions are neglected.

B. *Deep Learning Approaches*

In the last decade, most of the systems are developed on manually hand-crafted features but today those approaches have been exchanged with most recent deep neural networks approaches. For instance, the study conducted by Karatzas et al. [9] focuses on selective search approach along with deep neural networks to detect textual regions in scenery images. Gupta et al. [28] used YOLO architecture [10] to develop text detection model following fully convolutional DNN to localize text candidates. Output as textual regions of these systems is given as input to the DNNs for text recognition.

The work in Goodfellow et al. [29] focused text recognition model for house numbers. It was further improved by Jaderberg et al. [30] for every type of text recognition. In this system, single convolutional neural network is used that takes textual regions as input and perform text recognition (string text available in image). Complete end to end system proposed by Bissacco et al. [31] performs both text detection and recognition but text detection using traditional approached discussed above, manually hand-crafted features and then text candidates are binarized and provided as input to Deep FCNN that classifies each character region.

The work of Minghui et al. [4] provides word spotting and recognition end to end framework which is fast and accurate with single Deep Neural Network DNN named TextBoxes based on fully convolutional network FCN (LeCun et al. 1998). It outputs co-ordinates of text bounding boxes by determining text presence. Finally, aggregation of all boxes is the output using non-maximum suppression process. TextBoxes is trained on SynthText for 50k iterations and tune it up on ICDAR 2013 dataset for 2k iterations and finally ICDAR 2011 dataset is provided for test set. It outperforms on test set but failure in multi-oriented text-based images. Several systems for text detection and recognition using DNN are proposed by Jaderberg [30], [32].

The work in [32] developed bounding box regression CNN model for text detection and CNN model which performs classification based on textual regions as input, but it is limited to one single language as it classifies across pre-defined dictionary. In the study [30], sliding window approach is given for text detection and then CNN is used as sliding on textual regions in image. This CNN uses weight sharing with CNN for text detection. Work proposed by He et al. [33] uses both CNN and Recurrent Neural Network RNN. First, it creates slices for text candidates by using sliding window approach. Later, given input to text recognition CNN, this CNN produces features which are then forwarded to RNN to predict characters.

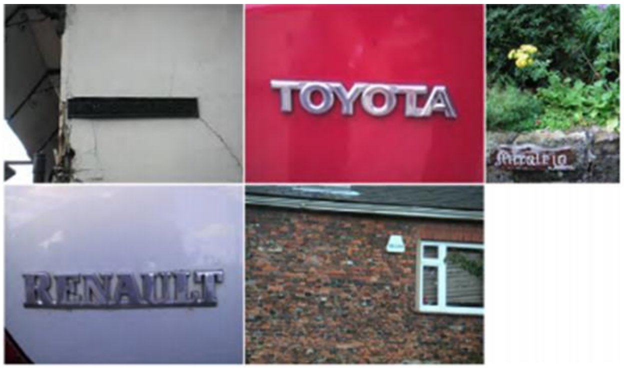


Fig. 2. Some Failures in Yuanwang et al. [16].

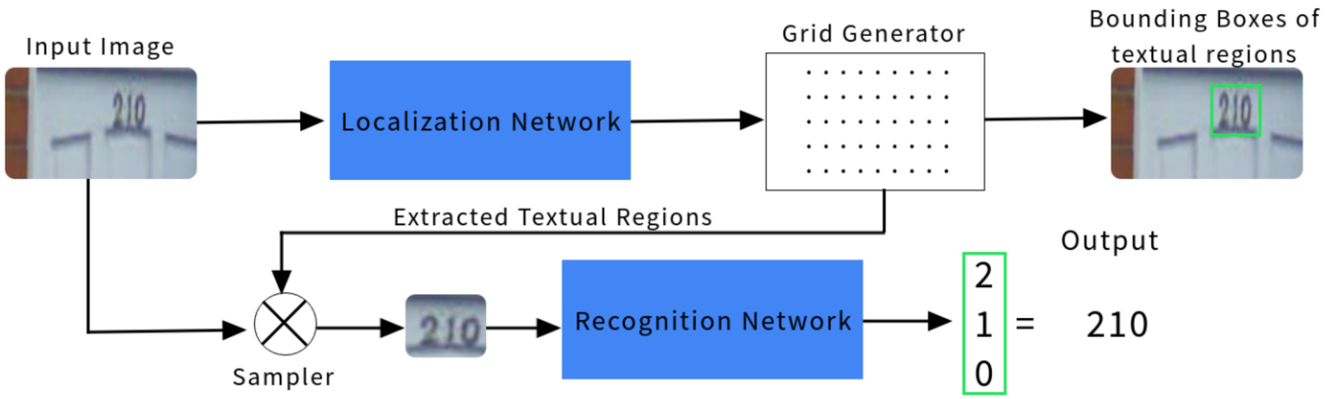


Fig. 3. **STN**-OCR Method for Text Detection and Recognition.

Hui et al. [6] presents method for text detection for horizontal-text in which major components including connected components extraction, character linking. Adaptive color reduction scheme is designed in this paper for CCs. Adjacent character model is also developed for connecting character which is trained through extreme machine learning. At the end, CNN with ELM is used for verification of text and non-text regions. Moreover, the method proposed in Yang et al. [5] is to detect and locate text in images by using two classifiers and non-text regions are cut by using local recursive search algorithm, and CNN is used to verify text candidates. Evaluation of proposed method is done through the ICDAR datasets such as ICDAR 2011, ICDAR 2013, ICDAR 2015. ICDAR datasets are benchmark used for specially for text detection from scene images. This method performed better on ICDAR datasets but multi-oriented texts are not detected.

The previous studies clearly show that there is still need of contribution like observing literature review [34], [35], most of the work is performed in text detection and recognition is based on single-orientation that is horizontal based and text detection in multi-orientation and multi-language is still very challenging task. Conferences such as International Conference on Document Analysis conferences and Recognition (ICDAR) or International Conference on Computer Vision ICCV are still held to find out latest research in this field. Multi-orientation and multi-language text identification are areas which needs to be explored.

Keeping the related studies in view, in this study, the system constituted based on the sliding window approach but with little changes. For instance, choice of sliding windows is not manually engineered but automatically learned by the model. Lastly, Spatial Transformer Network (STN) [36] is used as main building block for text detection.

III. Methodology

STN-OCR model behaves like a human, it will start reading line by line in sequential manner and read each character step by step. Most of the recent systems for scene text detection and recognition do not follow this human approach of reading text. These systems perform operations on complete image and extract all information at once. In this study, human-based approach is followed to find and localize textual regions sequentially in images and then recognize those localized textual regions. In this regard, Deep Neural Network (DNN) model is developed which is comprised of two stages: 1) text detection and 2) text recognition. This section will focus on attention mechanism used in text detection stage and complete structure of methodology for STN-OCR [3].

A. *Text Detection with Spatial Transformers*

This study has used Jaderberg et al. [36] proposed method which is Spatial Transformer, a learnable module for Deep Neural Networks that receives some input $I \in R^{H*W*C}$, performs some spatial transformations to input feature map I and then produces an output feature map *O*. There are three main parts for this spatial transformation. Localization network is the first part which computes function *f_loc*, predicting the parameters Θ of spatial transformation. The second part is used to create a sample grid based on predicted parameters Θ as input. It maps input features from predicted parameters on output feature map, in this part the sampling grid is generated, and that grid is provided to third part as input to learnable interpolation method and finally outputs transformed feature map *O*. Further in this section, we will describe each part in detail.

1) Localization network: In this part, feature map $\mathrm{I} \in \mathrm{R}^{\mathrm{H*W*C}}$ with height H, width W and C Channels is given as input to localization network and produces output predicted parameters Θ spatial transformation to be performed. In this part, this study's system predicts N two-dimensional transformation matrices $\mathrm{M}_{\theta}^{\mathrm{n}}$, where M is a matrix and $\mathrm{n} \in \{0,1,2, \dots, \mathrm{N}-1\}$.

Localization network will find and localize N number of characters, words or text lines. For achieving oriented text detection, network will be based on affine transformation matrices that will apply transformations including rotations, translations, skew and zoom to the input feature map *I*, in this regard this system learns to adopt and produces features based on text rotation, translation and zoom.

In STN-OCR, feed-forward CNN along with RNN is used to produce N affine transformation matrices M_{θ}^{n}. The CNN model ResNet-50 [33] is used in this localization network. Using this network structure, it is observed that system's performance is better than other structures like VGGNet [37] etc. It solves the problem of vanishing gradient and preserve better accuracy as in other network structure system's accuracy is not saturated. Furthermore, this study also used Batch

Normalization just for experiments and then use RNN in this part. RNN used here is Bi-directional LSTM. Prediction of affine transformation matrices is done through hidden states h_n, hidden states are basically generated by BLSTM.

- Localization Network Configuration

In localization network, residual neural network is used which is also known as ResNet architecture [38]. As this study is based on two stages so in this localization stage the images will be fed to the network where network will localize textual part. First layer of network will perform 3x3 convolution with 32 filters, second layer will perform same convolution with 48 filters and third layer with 48 filters. After each convolution layer, Batch Normalization [39] is performed followed by average pooling of 2x2 and stride 2. ReLU is used as activation function in each layer. After each layer, two residual layers are used with 3 x 3 convolution, each followed by Batch Normalization. After last residual layer, performed average pooling layer of 5 x 5 followed by BLSTM with 256 neurons. After above the model, sampling grid is generated where bounding boxes (BBoxes) are extracted for textual parts. BBoxes are extracted only for textual part as depicted in Fig. 4.

2) Generation of GRID: In this part, the system uses grid $G_0$with co-ordinates x_{w_0}, y_{h_0} along with affine transformation matrices produces N grids of input feature map *I*. During this step, *N* output grids are generated containing bounding boxes B-Boxes of textual regions localized by the network.

3) Image sampling: In the second part, grid generator produced *N* sampling grids, now they are used to sample values of feature map I at their respective coordinates for each $n \in N$. Logically, these points will not lie with exact grid values in feature map *I*. So, this study has used bi-linear sampling that selects nearest neighbors' points.

In Fig. 3, working of grid generator and image sampler are shown. After *N* output grids are produced by grid generator, these *N* grids are fed to image sampler which selects images pixels at that location by using those sampling grids. This system automatically generates Bboxes by generated sampling grids vertices. Hence, combining these three-parts localization network, generation of grid and image sampling formulates Spatial Transformer that can be used generally in every part of Deep Neural Network. Spatial Transformer is first step in this system.

B. *Text Recognition*

Text detection stage returns *N* textual regions which are extracted from the input image. In this text recognition stage, each *N* regions are handled independently of each other. Processing of *N* regions is done by CNN.

Variant of ResNet is used in this CNN too because it was observed that ResNet producing better results in text recognition system. Text detection needs to obtain strong gradients from text recognition stage. Basically, in this stage, probability distribution over label space is predicted. Softmax classifiers are used to predict probability distribution.

$$x^n = O^n$$

$$y_t^n = softmax(f_{rec}(x^n))$$

$$y^n = \sum_{t=1}^{T} y_t^n$$

After applying convolution feature extractor, we obtain the result $f_{rec}\ (x)$.

Recognition Network Configuration

Configuration of recognition network is same as in localization except convolution filters. This network contains total three convolutional layers having filters of 32, 64 and 128.

C. *Training Network*

ICDAR 2015 [40] is used to train the network, the training input set X used for training the network/model comprised of images and separate text file for each image. Each file contains coordinates *x1, y1, x2, y2, x3, y3, x4, y4,* label for words in each image where x1, y1 are top-left coordinates, x2, y2 are top-right coordinates, x3, y3 are bottom-right coordinates and x4 and y4 are bottom-left coordinates. Label is not used in the first stage which is text detection because at this stage the model is only learning localization and finding text candidates, but in next stage – text recognition the model uses label for recognition.

Text detection learns to find and localize text regions by using error gradients by calculating loss of text labels prediction. Text detection is performed with some pre-training steps because we observed initially that model does not combine text if there are multi-line texts within image. Optimization algorithm has great impact on network while training the model. It is observed that Stochastic Gradient Descent (SGD) is better on simpler tasks during pre-training the network and after pre-training the network using SGD, Adam optimizer [41] is applied for improving the already trained network. In text detection stage, the learning rate is kept constant in the first stage for longer period. This is resulting in finding and better localizing textual regions. Therefore, SGD is used and it works better for this. Text recognition stage further starts learning to recognize already predicted text regions from previous stage.

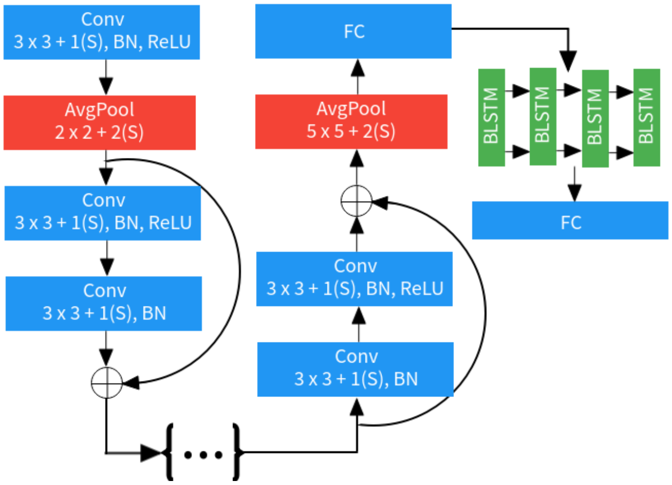


Fig. 4. Localization Network.

IV. Results and Discussions

This section discusses about the experiments, results and discussions which are achieved by using the study model 's architecture.In this study, two benchmarks datasets ICDAR 2015 [40] and Street View House Numbers (SVHN) are used. Above two datasets are discussed below.

- ICDAR 2015 Dataset

ICDAR 2015 dataset is used in Robust Reading Competition, containing total 1500 images with over 10k annotations. 1000 images are used for training and remaining 500 images are used for testing. Along with that annotations for images include text regions. ICDAR 2015 basically is used for three tasks: text localization, word recognition and end-to-end recognition. For text localization, it provides bound boxes(Bboxes) of text for each image. Bboxes are in separate file, containing BBoxes for each image separated by line. For Word recognition, it provides that word too along with BBoxes. See Fig. 5 (Taken from ICDAR Official Site).

The text file contains BBoxes and word for all images separated by line in format:

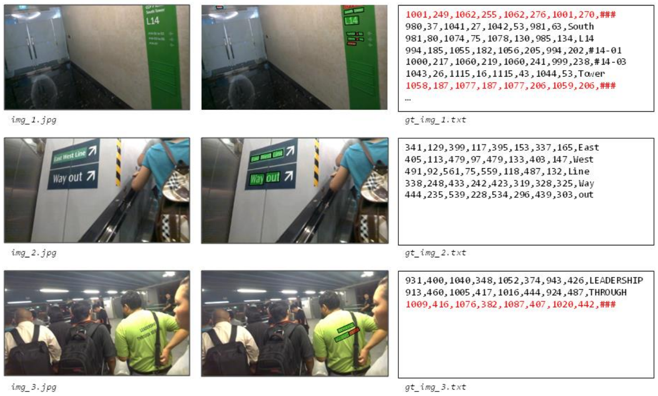
x1, y1, x2, y2, x3, y3, x4, y4, transcription

- SVHN Dataset

Street View House Numbers (SVHN) is benchmark dataset containing low resolution images and requiring low data processing and formatting. It can be said that it is like MNIST dataset [42]. But this contains a lot of variety of image including low resolutions, blurred images as this dataset has been developed from house numbers in Google Street View Images. It comes into two formats, one is like MNIST, cropped digits images and second in complete house door images along with bounding boxes for digits. Dataset contains too many images, 73257 digits for training and 26032 for testing.

- Expeiments on Datasets

The first dataset used for experimentation is ICDAR 2015. It is most challenging dataset because it consists of several images including different background, noise, cluttered, dot matrix fonts, blurry images and low resolution etc. Results are shown in Fig. 6.



```
1001,249,1062,255,1062,276,1001,270,###
980,37,1041,27,1042,53,981,63,South
981,80,1074,75,1078,130,985,134,L14
994,185,1055,182,1056,205,994,202,#14-01
1000,217,1060,219,1060,241,999,238,#14-03
1043,26,1115,16,1115,43,1044,53,Tower
1058,187,1077,187,1077,206,1059,206,###
…
```

```
341,129,399,117,395,153,337,165,East
405,113,479,97,479,133,403,147,West
491,92,561,75,559,118,487,132,Line
338,248,433,242,423,319,328,325,Way
444,235,539,228,534,296,439,303,out
```

```
931,400,1040,348,1052,374,943,426,LEADERSHIP
913,460,1005,417,1016,444,924,487,THROUGH
1009,416,1076,382,1087,407,1020,442,###
```

Fig. 5. (a) Original Image (b) Visualization of Text (c) BBoxes Text File.

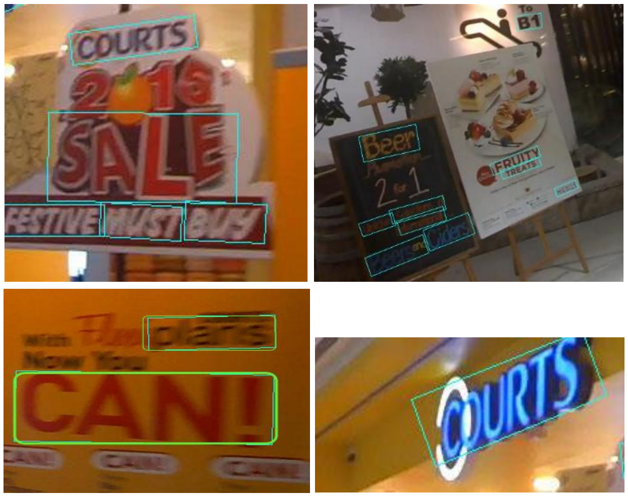


Fig. 6. Results on ICDAR 2015 Incidental Scene Text.

Furthermore, the model architecture of STN-OCR is evaluated on ICDAR 2015 which outperforms and produce good results in multi-oriented text, low resolution, dot-matrix fonts etc as shown in Fig. 7. It includes 1500 different variety of images including noise, low resolution, multi-oriented text, etc.

STN-OCR method outperforms by achieving 65.2%, 78.53% and 71.86% of Recall, Precision and H-mean respectively. Fig. 8 keeps the comparision side by side.

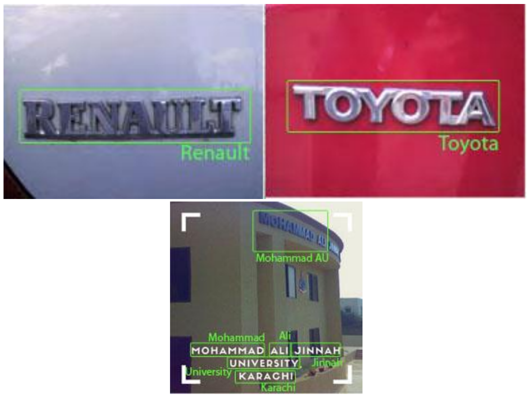


Fig. 7. Results of this System on Failure Cases in Existing Work.

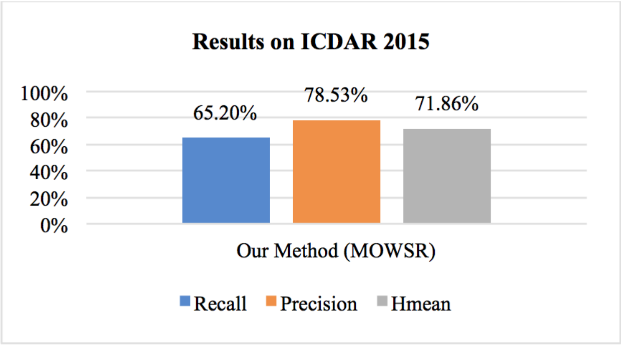


Fig. 8. STN-OCR performance

Table I clearly shows that proposed method has achieved better results comparing others.

SVHN is the second dataset on which the evaluatation of the network architecture is performed to prove this model can work on real data. In SVHN, house numbers are containing noise too. After experiments on SVHN dataset, it was observed that this study network architecture works on SVHN house numbers by finding, localizing and recognizing house numbers on sampling grid. For achieving best results, the study model was trained from very beginning by initializing random weights but only first stage that is localization network initialized with weights from already trained network. In this regard, localization network stage tends to produce better results. Table II shows accuracies on SVHN dataset while text recognition on real data which is house numbers.

TABLE I. RESULTS OF DIFFERENT APPROACHES ON ICDAR 2015 [40]

Method	**Recall**	**Precision**	**Hmean**
Tencent Youtu	60.42%	79.83%	68.79%
Baidu VIS v2	66.59%	69.95%	68.23%
SRC-B-Machine Learning Lab	61.72%	74.62%	67.56%
Baidu VIS	63.02%	71.37%	66.94%
HoText_v1	63.46%	68.36%	65.82%
FOTS	53.20%	84.61%	65.33%
STN-OCR	**65.20%**	**78.53%**	**71.86%**

TABLE II. RESULTS ON SVHN DATASET

Method	**Accuracy**
MaxoutCNN [28]	96
ST-CNN [37]	96.3
STN-OCR	**97.5**

After ICDAR 2015, this system achieved better results with 97.5% accuracy on SVHN. Some results which are not handled in existing work [16] were already discussed in literature, but the study model works good on those images. This sutdy is experimented on Google's K80 GPU, 12GB RAM for testing the model, and it produces results within 2-3 seconds per image with color background.

V. CONCLUSION

In this study, end to end text detection and recognition model (STN-OCR) using single DNN is applied on latest benchmark datasets such as ICDAR2015. This system consists of two stages: text detection and text recognition. Text detection model finds and localizes text regions in image and then output of this stage is input of text recognition network which recognizes text regions in image. The main purpose was to detect multi-oriented text and it was achieved with better results. The study clearly shows that this model achieves 97.5% accuracy on SVHN and performed better on ICDAR 2015 than state-of-art methods. This model still is limited to combine words to make complete line/sentence. Moreover, future work includes to implement this model on other local/famous languages (i.e. Urdu/Hindi) and adjusting geometry design for finding directly curved texts.

REFERENCES

[1] Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 355–372.

[2] A. A. Panchal, S. Varde, and M. S. Panse, "Character detection and recognition system for visually impaired people," in 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016, pp. 1492–1496.

[3] C. Bartz, H. Yang, and C. Meinel, "See: Towards semi-supervised end-to-end scene text recognition," in 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 2018.

[4] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[5] Y. Zheng, Q. Li, J. Liu, H. Liu, G. Li, and S. Zhang, "A cascaded method for text detection in natural scene images," Neurocomputing, vol. 238, pp. 307–315, 2017.

[6] H. Wu, B. Zou, Y. Zhao, and J. Guo, "Scene text detection using adaptive color reduction, adjacent character model and hybrid verification strategy," Vis. Comput., vol. 33, no. 1, pp. 113–126, 2017.

[7] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "Textfield: Learning a deep direction field for irregular scene text detection," IEEE Trans. Image Process., 2019.

[8] A. Sain, A. K. Bhunia, P. P. Roy, and U. Pal, "Multi-oriented text detection and verification in video frames and scene images," Neurocomputing, vol. 275, pp. 1531–1549, 2018.

[9] L. Gómez and D. Karatzas, "Textproposals: a text-specific selective search algorithm for word spotting in the wild," Pattern Recognit., vol. 70, pp. 60–74, 2017.

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

[11] S. Khan, D.-H. Lee, M. A. Khan, A. R. Gilal, and G. Mujtaba, "Efficient Edge-Based Image Interpolation Method Using Neighboring Slope Information," IEEE Access, vol. 7, pp. 133539–133548, 2019.

[12] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in Asian Conference on Computer Vision, 2010, pp. 770–783.

[13] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3538–3545.

[14] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 97–104.

[15] A. Alshanqiti, A. Bajnaid, A. Rehman, S. Aljasir, A. Alsughayyir, and S. Albouq, "Intelligent Parallel Mixed Method Approach for Characterising Viral YouTube Videos in Saudi Arabia," Int. J. Adv. Comput. Sci. Appl., 2020.

[16] Y. Wei, Z. Zhang, W. Shen, D. Zeng, M. Fang, and S. Zhou, "Text detection in scene images based on exhaustive segmentation," Signal Process. Image Commun., vol. 50, pp. 1–8, 2017.

[17] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in 2011 18th IEEE International Conference on Image Processing, 2011, pp. 2609–2612.

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

[19] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," IEEE Trans. Image Process., vol. 20, no. 9, pp. 2594–2605, 2011.

[20] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 5, pp. 970–983, 2013.

[21] L. Neumann and J. Matas, “Text localization in real-world images using efficiently pruned exhaustive search,” in 2011 International Conference on Document Analysis and Recognition, 2011, pp. 687–691.

[22] T. Quack, Large-scale mining and retrieval of visual data in a multimodal context, vol. 53. ETH Zurich, 2008.

[23] M. Anthimopoulos, B. Gatos, and I. Pratikakis, “Detection of artificial and scene text in images and video frames,” Pattern Anal. Appl., vol. 16, no. 3, pp. 431–446, 2013.

[24] I. Posner, P. Corke, and P. Newman, “Using text-spotting to query the world,” in 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010, pp. 3181–3186.

[25] A. Mishra, K. Alahari, and C. V Jawahar, “Scene text recognition using higher order language priors,” 2012.

[26] C. Yao, X. Bai, B. Shi, and W. Liu, “Strokelets: A learned multi-scale representation for scene text recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4042–4049.

[27] K. Wang and S. Belongie, “Word spotting in the wild,” in European Conference on Computer Vision, 2010, pp. 591–604.

[28] A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic data for text localisation in natural images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2315–2324.

[29] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, “Multi-digit number recognition from street view imagery using deep convolutional neural networks,” arXiv Prepr. arXiv1312.6082, 2013.

[30] M. Jaderberg, A. Vedaldi, and A. Zisserman, “Deep features for text spotting,” in European conference on computer vision, 2014, pp. 512–528.

[31] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, “Photoocr: Reading text in uncontrolled conditions,” in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 785–792.

[32] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Reading text in the wild with convolutional neural networks,” Int. J. Comput. Vis., vol. 116, no. 1, pp. 1–20, 2016.

[33] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, “Reading scene text in deep convolutional sequences,” in Thirtieth AAAI conference on artificial intelligence, 2016.

[34] A. R. Gilal, J. Jaafar, M. Omar, S. Basri, and A. Waqas, “A Rule-Based Model for Software Development Team Composition: Team Leader Role with Personality Types and Gender Classification,” Inf. Softw. Technol., vol. 74, pp. 105–113, 2016.

[35] A. R. Gilal, J. Jaafar, L. F. Capretz, M. Omar, S. Basri, and I. A. Aziz, “Finding an effective classification technique to develop a software team composition model,” J. Softw. Evol. Process, vol. 30, no. 1, pp. 1–12, 2018.

[36] M. Jaderberg, K. Simonyan, A. Zisserman, and others, “Spatial transformer networks,” in Advances in neural information processing systems, 2015, pp. 2017–2025.

[37] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv Prepr. arXiv1409.1556, 2014.

[38] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[39] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv Prepr. arXiv1502.03167, 2015.

[40] D. Karatzas et al., “ICDAR 2015 competition on robust reading,” in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1156–1160.

[41] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv Prepr. arXiv1412.6980, 2014.

[42] Y. LeCun, “The MNIST database of handwritten digits,” http://yann. lecun. com/exdb/mnist/, 1998.