Natural Scene Text Detection and Localization using EAST model with ResNet-50 network

Manasi Shrivastava [1], Ranjeet Kumar Singh [2]

[1] B. Tech. Student, Dept. of Computer Science and Engineering, Madhav Institute of Technology and Science, Gwalior, India
[1]manasishrivastava362@gmail.com
[2] Assistant Professor, Dept. of Computer Science and Engineering, Madhav Institute of Technology and Science, Gwalior, India
[2]2014rsca002@gmail.com

Abstract. The ever-decreasing reliance on traditional and error-free methods of acquiring documents and images has long since proven challenging for image processing methods like Optical Character Recognition to produce successful detection rates for printed characters in suboptimal conditions like bad light, blur, incomplete/incoherent characters in natural scenes. These raise difficulties in image processing and many real-life problems concerning data extraction, script recognition, security and surveillance. OCR has also become a day-to-day necessity (E.g., Google Lens) and it needs to provide utmost precision in its outputs. In this paper, the process of OCR is studied and implemented on natural scene text images using deep learning model EAST (with ResNet-50 network), Tesseract, and Computer Vision libraries in Python to better understand and describe the pipeline involved in training simple yet robust OCR systems to try and address a major concern of recognition accuracy due to an inherent degree of vagueness and imprecision present in real-world data. Experimental results include a comparative study of qualitative and quantitative results in text localization in natural scene images as well as accuracy and speed comparisons of various text-detection methods on the dataset used.

Keywords: Image Processing, OCR, natural scenes, EAST, Tesseract, Computer Vision, Deep Learning

1 Introduction

OCR (Optical Character Recognition) is a method of converting handwritten, typewritten, or printed text characters into machine-readable text. It is one of the most popular areas of research in pattern recognition and deep learning. An actively studied and researched topic in industry and academia, it has immense application potential too. Since its advent in the 1930s, OCR has mainly been used in digitising old documents, working on translations and automated data extraction for large organisations. But over time as deep learning tools have expanded, researchers have been working to develop sophisticated and robust OCR systems to make them as intelligent and fool-proof as the human eye.

Nowadays, OCR is being used in every possible industrial undertaking, from self-driving cars and medical scans to a simple google image search. But with its increasing popularity and ease of availability have come some pressing issues.

Traditional scanners have mostly been abandoned for easy-to-use handheld imaging devices (HIDs) like mobile phone cameras, which take less than optimal images resulting in inaccurate outputs.

Natural scenes, also called real-world scenes, are the images that represent natural environment, i.e., complex background with bushes, trees, fields, sky, etc. Scene text detection, when done expertly keeping image processing in mind with good equipment is easy for most models to detect scene text; but the scenes taken from HIDs consist of outdoor and non-planar/non paper objects, unknown layout, natural scene text and objects in distance. Some of the conditions that these scenes bring are raw sensor image and sensor noise, incongruent viewing angles, blur and sparse/thick lighting, inapt resolution and aliasing; which make image processing harder. [2]

To mitigate these challenges and to build better constrained and controlled environment for image processing, EAST (Efficient Accurate Scene Text Detector) was proposed by Zhao *et al* [1] in 2017. EAST is a deep learning model based on novel architecture and training patterns. The system is equipped with a light-weight, single neural network and can run both on images and videos at near run-time at 13.2*fps* (Frames Per Second) on 720p images with an 0.7280 F-score on the ICDAR 2015 dataset.

EAST's scene text detection model contains two stages: a Fully Convolutional Network (FCN) stage and a Non-Max Suppression (NMS) merging stage. The motive behind NMS merging is to produce words for all sizes of text regions. FCN stage, in this paper, utilises the ResNet-50 backbone for feature extraction. The network outputs the geometries in the form of rotating boxes. In addition, score maps are created to indicate the likelihood of positive text locations, allowing for easier feature extraction and merging. A region-based loss function called Dice Loss was used to optimally cover even the edge cases. In this paper, the model is trained using AdamW, an Adam optimizer variant with an improved weight decay implementation in order to prevent overfitting. During implementation of the model, the score map is reverted back to their bounding boxes and thresholding is done to discard the boxes with low probability of positive text regions. Non-Max Suppression is then applied to the remaining boxes to yield final results.

2 Background

2.1 Workflow of the model

The general flow-work diagram of the detection and recognition process from the input image in this model is shown below:

Fig. 1. Flow-work

The detector detects the words (text regions, global and local) in the image and makes the bounding boxes, and the recognizer recognizes the detected words.

2.2 Pipeline

The pipeline of the EAST scene text detection model is shown below, it eliminates the majority of intermediate steps including transformation of the image:

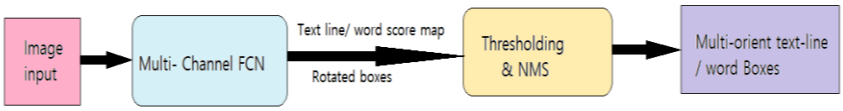

Fig. 2. Pipeline of the EAST model

3. Model Framework Overview

Figure 3 represents the complete framework of the EAST model. The feature extraction stem consists of the FCN, the feature merging stage involves thresholding and Non-Max Suppression.

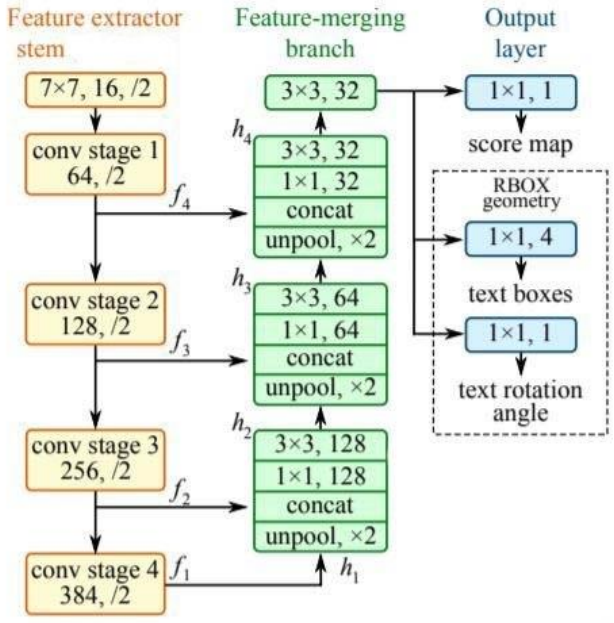

Fig. 3. Structure of the Fully Convolutional Network

3.1 Fully Convolutional Network stage (FCN)

FCN is an end-to-end trained, pixels-to-pixels network that predicts dense outputs from arbitrary-sized input images. This allows us to produce a hierarchical order of features from the objects to further extract the information we need from the redundant layers. It helps in zeroing in on the text from natural scenes without the irrelevant features.[3]

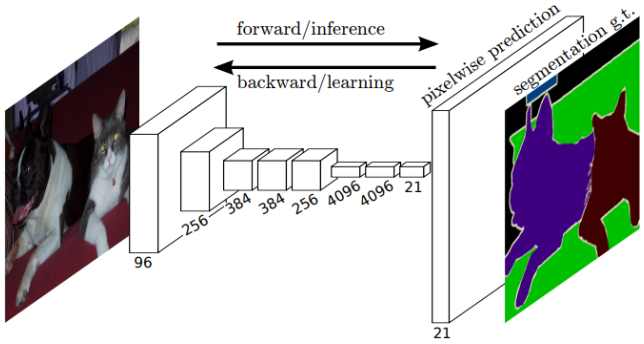

Fig. 4. FCN can efficiently learn pixelwise prediction in semantic segmentation by forward and backward inference.

In this stage, ResNet-50 pre-trained on Imagenet dataset is used as feature extractor. Four layers of feature maps can be obtained from this network: f1, f2, f3, and f4 as is shown in figure 3.

One of FCN's expected channels is a score map with pixel values between [0,1]. The remaining channels are geometries that surround the word from the perspective of each pixel. The score represents the degree of certainty in the geometry shape at the predicted at the same location.

- RBOX

The network used in the paper outputs geometries in the form of RBOX. RBOX uses a four-channel (top and left coordinate, height and width) axis-aligned bounding boxes with a channel rotation angle θ.

Table. 1. Channels and Description of the rotating box geometry

Geometry	Channels	Description
RBOX	5	G = {R, θ}

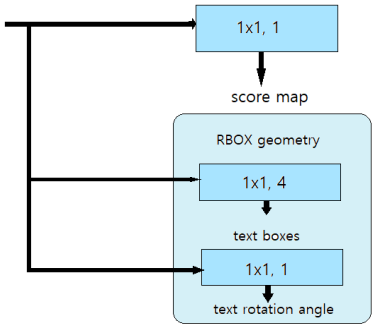

Fig. 5. RBOX geometry and Score Maps

- Score-Map

The score maps used in deep leaning for object detection are complex-layered (convoluted) feature maps that recognize parts of an object. In the model, score-map generation is done by shrinking the bounding box and all the pixels inside it represent the positive score map. These scores depict the probabilities for positive text regions.

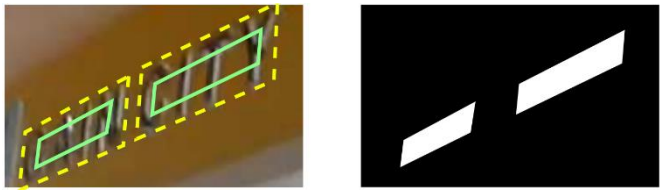

Fig. 6. Text Score Map

3.3 Non-Max Suppression stage

NMS is an object detection technique with a primary purpose of identifying and discarding the redundant edges in the bounding boxes that are below a given probability bound. In the paper, the process is repeated amongst a large number of overlapping bounding boxes in the input image until the maximum probable positive text region is selected.

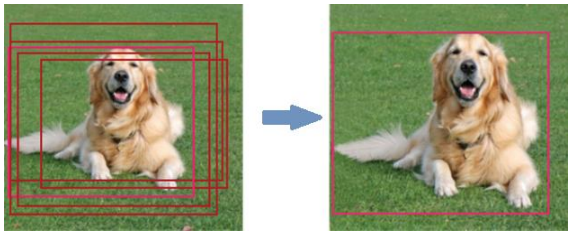

Fig. 7. Predictions before and after applying NMS

4. Model Implementation and Design

The algorithms of the generation function, loss function and LANMS are described in brief in this section.

4.1 Generator Function

In order to prepare the data to be fed into the model, a generator function has been built which produces the input of the model (Input array) with a score map output from the multichannel Fully Convolutional Network.

To generate a score map, each boundary of the bounding box is shrunk edgewise by moving its vertices inwards by a reference length.

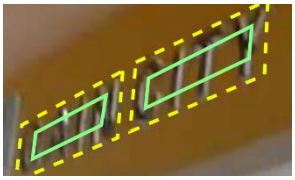

Fig. 8. Text quadrangle (yellow dashed) and the shrunk quadrangle (green solid)

4.2 Loss Function

Loss functions are mathematical expressions used in deep learning algorithms to optimize and learn the objective. In this paper, a region-based loss function called Dice Loss has been used to calculate the similarity between two images during image processing [4].

Dice loss (a function of Intersection and Union over foreground pixels) maximizes the intersection area over foreground resulting in losing most of the unnecessary pixels in the background, thus solves the data imbalance problem in the foreground and background.

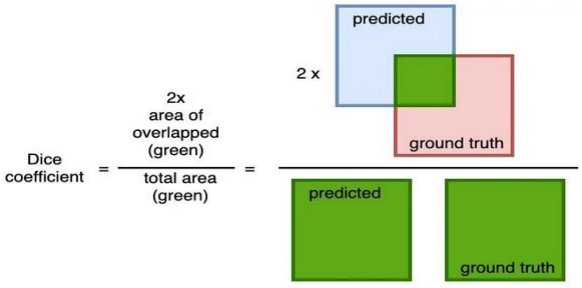

Fig. 9. Dice Coefficient overlapping of predicted and ground truth information

4.3 Locally Aware Non-Max Suppression

First proposed in [1] EAST test detection, LANMS reduces the calculation required to compute thousands of geometric bounding boxes, as is the case with natural scene images.
In this paper, the cluster of boxes obtained as output in FCN is combined with their corresponding thresholds. Thresholding is applied to remove low confidence boxes. The boxes with greater thresholds are combined by weighted merge after undergoing needed iterations. After this, standard NMS is applied to the remaining boxes.

5. Experiment

5.1 Dataset

The input data have been collected from the incidental scene-text dataset ICDAR-2015 provided by International Conference of Document Analysis and Recognition. The dataset consists of training set with of images and testing set of 500 images with ground truth information for each set.

These images are obtained from wearable cameras. They are multi-oriented with small and blurred text regions in the English language.

Fig. 10. Examples of ICDAR-2015 testing set images

5.2 Network

In this paper, a pretrained ResNet-50 model has been utilized to extract features from input images from the ICDAR-15 database. It is a Convolutional Neural Network that is 50 layers deep. It drops the redundant features of the text from images, i.e., curvy shaped fonts, size colour, and background of the scene.

6. Results

6.1 Qualitative Results

On conducting text localization and detection using the EAST model with ResNet-50 as feature extractor, following results were produced. Figure 11 shows an incidental scene from the testing set of ICDAR-15 and its predicted image with the bounding boxes respectively.

Fig. 11. Original Image and Predicted Image

The primary goal of the EAST model, which used ResNet-50 as a feature detector, was to improve the extracted text in order to eventually increase character recognition rates in natural scenes. A few qualitative results of the model are depicted in figure 12.

Fig. 12. Some exemplary results of the EAST text detection model on images from the ICDAR-2015 testing set

6.2 Quantitative Results

Table. 2. State-of-the-art Results of Text Localization by different algorithms on the ICDAR 2015 dataset

Method	**Recall (%)**	**Precision (%)**	**F-measure (%)**
EAST+ResNet-50 RBOX [Our Project]	**77.32**	**84.66**	**80.83**

EAST+PVANET2x RBOX [1]	73.47	83.57	78.20
EAST+ VGG16 RBOX [1]	72.75	80.46	76.41
Yao et al. [6]	58.69	72.26	64.77
Tian et al. [7]	51.56	74.22	60.85
Zhang et al. [8]	43.09	70.81	53.58
CNN MSER [9]	34.42	34.71	34.57

Table.3. Devices used, and the time and frames per second of different methods on the ICDAR 2014 dataset

Method	**Device**	**Time (ms)**	**FPS**
EAST + ResNet-50 RBOX [Our Project]	Radeon 540	**64.0**	**15.6**
EAST+PVANET2x RBOX [1]	Titan X	73.8	13.2
EAST+ VGG16 RBOX [1]	Titan X	150.9	6.52
Yao et al. [6]	K40M	420	1.61
Tian et al. [7]	GPU	130	7.14
Zhang et al. [8]	Titan X	2100	0.476

7. Result Analysis

The primary goal of this paper was that EAST model, when reimplemented with ResNet-50 as a feature detector, shows significant improvements in the extracted text which eventually increases character recognition rates in natural scenes. The experiments conducted on a server using a single Radeon 540 graphic card and an Intel i5-8250U @ 1.60 GHz CPU. As can be seen from the tables, the model variation used in the paper gives the most accurate as well as less time consuming results among all the parameters compared to previous models and EAST with PVANET.

Recall = TruePositives / (TruePositives + FalseNegatives)

and,

Precision = TruePositives / (TruePositives + FalsePositives)

Recall and precision are indicators of a machine learning model's performance. They show how accurate a model is when tested.

From the Tables, it is clear that the model variant in this project has both the best recall (0.7732) and precision (0.8466) resulting in a f-score of 0.8083. Similarly, it can seen that the model is less time-consuming with a better frames per second rate, 15.6*fps* on 720p images as compared to 13.2*fps* of the original EAST model.

8. Conclusion

Deep learning based Optical Character Recognition methods perform better for unstructured data like natural scenes. These focus on discarding complex background features by feature extraction performed during processing. The EAST model follows: a fully Convolutional Network directly generates pixel-based word predictions for bounding boxes that contain the text regions. The next stage involves feature merging where the feature maps are un-pooled and concatenated with a single feature map in each merging stage using Locally Aware Non-Max Suppression algorithm. Dice loss function is incorporated to predict rotated rectangles for the text regions. Comparison between OCR system performance on structured document images and unstructured natural images was presented for some pre-processing techniques. Experimental localization and detection of images from the dataset were also presented to depict the accuracy and efficiency of the entire EAST system with ResNet-50 network. The model presented in this paper has a F-measure of 80.83% on the ICDAR 2015 dataset. The model also has better frames per second rate than the original EAST model with 15.6 FPS and overall faster speed in prediction by network and post-processing.

9. Related work

Preliminary research was done by learning about the various pre-processing techniques involved in character recognition as discussed by Alginahi [5]. Histogram plotting and Line Detection have been conducted on the document images through that knowledge. Mancas-Thillou [2] has stated that the main challenge is to design a text detection and recognition system that can manage as much variation as possible in daily life, such as variable targets with unknown layouts, scene text, multiple character fonts and sizes, and variation in imaging conditions such as uneven lighting, shadowing, and aliasing. As text detection is the first step in a text understanding system, it is essential in the entire process of character recognition in natural scenes. The construction of features to identify text from backgrounds is at the heart of text detection. Both classic and deep learning methods are inefficient and time-consuming, yielding subpar outcomes. To address such issues, Zhou *et al* [1]

proposed a text detection model that eliminates unnecessary intermediate components and steps. The resulting system can run both on images and videos at near run-time on 720p with speed 13.2FPS (Frames Per Second) with an 0.7280 F-score on the ICDAR 2015 dataset, with PVAnet as feature extractor. Fully Convolutional neural network which a key component of the EAST framework has been adapted and discussed in detail in [3] by Long *et al* for the pre-processing stage of feature extraction. The process of semantic segmentation (labelling each pixel in an image) is done by using loss functions, specifically the Dice Coefficient which is reviewed along with others by Jadon [4]. The post-processing algorithm used is non-Max suppression as discussed in detail [12] It selects the high probability text region and removes the low confidences bounding boxes by thresholding, resulting in better predictions and fast detections. The different models discussed briefly in table. 2 and table. 3 are based on Holistically-Nested Edge Detection (HED) method [6], a vertical convolutional approach [7], detection using a VGG 16-layer net and Character-Centroid FCN [8] and a MSER-based convolutional neural network approach [9].

10. References

[1] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang: "EAST: An Efficient and Accurate Scene Text Detector", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21-26 July 2017.

[2] Mancas-Thillou, Céline; Gosselin, Bernard: Natural Scene Text Understanding. Vision Systems: Segmentation and Pattern Recognition, June 2007, pp. 302-332.

[3] Long, Jonathan; Shelhamer, Evan; Darrell, Trevor: Fully Convolutional Networks for Semantic Segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), October 15th 2015.

[4] Jadon, Shruti: A Survey of Loss Function for Semantic Segmentation. IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 27-29 Oct. 2020.

[5] Alginahi, Yasser: "Preprocessing Techniques in Character Recognition" in M. Mori (Ed.) Character Recognition. London, United Kingdom, IntechOpen, 2010, August 17th, 2010, pp. 1-20.

[6] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao. "Scene text detection via holistic, multi-channel prediction", 2016.

[7] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, “Detecting text in natural image with connectionist text proposal network.”, In European Conference on Computer Vision, pages 56–72. Springer, 2016.

[8] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. “Multi-oriented text detection with fully convolutional networks”. In Proc. of CVPR, 2015.

[9] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny. ICDAR 2015 competition on robust reading. In Proc. of ICDAR, 2015.

[10] Niblack, W., “An introduction to image processing”, Prentice-Hall, 1986, pages 115-116

[11] J. Gllavata , R. Ewerth and B. Freisleben, “A Text Detection, Localization and Segmentation System for OCR in Images:, IEEE Sixth International Symposium on Multimedia Software Engineering, 2004.

[12] Jan Hosang, Rodrigo Benenson and Bernt Schiele, “Learning non-maximum suppression”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[13] L. Neumann and J. Matas, ‘‘A method for text localization and recognition in real-world images,’’ in Proc. 10th Asian Conf. Computer Vision. (ACCV). Berlin, Germany: Springer, 2010, pages 770–783

[14] Casey, R.G. & Lecolinet, E. (1996). “A survey of methods and strategies in character segmentation”, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 18, No. 7, pages 690-706

[15] Gllavata, J.; Ewerth, R. & Freisleben B. (2003). “Finding text in images via local thresholding”, Proceedings of IEEE Symposium on Signal Processing and Information Technology, pages 539-542

[16] Text Recognition Algorithm Independent Evaluation (TRAIT) Accessed: 2015-11-1.

[17] 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 Trans. Pattern Analysis and Machine Intelligence, 2016.

[18] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park. “PVANET: Deep but lightweight neural networks for real-time object detection”, 2016