Recognition of the Multioriented Text Based on Deep Learning

K. Priyadarsini, Senthil Kumar Janahan, S. Thirumal, P. Bindu, T. Ajith Bosco Raj, and Sankararao Majji

Abstract The development and use of systems for analyzing visuals, such as photos and videos, using benchmark datasets is a difficult but necessary undertaking. DNN and STN are employed in this study to solve the challenge at hand. The study's network design consists of a localization and recognition network. The localization network generates a sampling grid and locates and localizes text sections. In contrast, text areas will be entered into the recognition network, and this network will then learn to recognize text, including low resolution, curved, and multi-oriented text. Street View house numbers and the 2015 International Conference on Document Analysis and Recognition were used to gauge the system's performance for this study's findings (ICDAR). Using the STN-OCR model, we are able to outperform the literature.

Keywords Spatial transformer networks · Deep neural networks · Recognition · STN-OCR · Multi-oriented text etc

K. Priyadarsini
Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

S. K. Janahan (✉)
Department of CSE, Lovely Professional University, Phagwara, Punjab, India
e-mail: senthil.26610@lpu.co.in

S. Thirumal
Department of Computer Science and Engineering, Vels Institute of Science Technology and Advanced Studies, Chennai, India

P. Bindu
Department of Mathematics, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

T. A. B. Raj
Department of Electronics and Communication Engineering, PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India

S. Majji
Department of Electronics and Communication Engineering, GRIET, Hyderabad, India

I. J. Jacob et al. (eds.), *Expert Clouds and Applications*, Lecture Notes in Networks and Systems 444, https://doi.org/10.1007/978-981-19-2500-9_26

1 Introduction

Increasing demand for numerous computer vision jobs has pushed this community to focus on reading text in the wild (from scene photos). Despite substantial research in the last few years, finding text in uncontrolled contexts remains a difficult task [1, 2]. Even more challenging is recognizing text lines with random orientation, which takes into account a substantially greater number of hypotheses, which significantly expands the search field. In most cases, existing methods are able to recognize text that is either horizontal or close to horizontal. But when applied to multi-oriented text, the results of the recent ICDAR2015 competition for text detection show that there is still a considerable disparity.

Text has a fairly different appearance and shapes when compared to generic objects since it can be handled as a sequence-like object with unlimited lengths [3–5]. This has led to the widespread use of scene image identification systems based on sliding windows and related components. ICDAR2013 and ICDAR2015 contests saw state-of-the art performance from component-based approaches using Maximally Stable Extremal Regions (MSER) as the fundamental representations. An extremely resilient representation of character components was recently learned through the use of a convolution neural network (CNN). To localize a word or a line of text, clustering algorithms or some sort of heuristic approach is usually required for this. Directly hits text lines from crowded photos, taking advantage of their symmetry and self-similarity features [6]. Text detection appears to need the use of both character components and text regions.

2 Related Work

Natural picture text identification has piqued the curiosity of computer vision and document analysis professionals. Horizontal or near-horizontal-based text detection is the primary focus of most techniques. In order to create an end-to-end text recognition system, the first step is to locate word boundaries [7].

Here, we'll take a look at some of the best examples of multi-oriented text detection. Lu et al. [8]. were the first to look at real-world multi-oriented text detection. Conventional detection pipelines can be compared to those that use connected component extraction and text line orientation estimation. Kang, et al. [8] turned the text identification problem into a graph partitioning problem by treating each MSER component as a node in a network. It has been proposed by Wei et al. [9]. Yin and others use multi-stage clustering methods in order to recognize multi-oriented text [10]. For multi-oriented text, an SWT-based end-to-end system was proposed by Yao and his colleagues. The ICDAR2015 text detection competition just announced a hard benchmark for multi-oriented text identification, and numerous academics have presented their results on it [11].

3 Methodology

It reads line by line, character by character, just like a person would using the STN-OCR model. This human-like method to text analysis is no longer employed by text detection and recognition algorithms. An image is processed in its entirety, allowing these systems to retrieve all relevant information at once. Textual sections are found and localized progressively in photos using a human-based technique, and then recognized [12, 13]. Text detection and recognition are part of a Deep Neural Network (DNN) model that was created in this regard. This section focuses on the text detection stage's attention mechanism and the complete approach for STN-OCR (Fig. 1).

A. *Text Detection with Spatial Transformers*

Jaderberg et al. employed the Spatial Transformer, a Deep Neural Networks learnable module that receives input I, spatially modifies the input feature map, and then outputs an output feature map O. This shift in location is made up of three main components. The initial part of the localization network computes the function f_ loc, which predicts the spatial transformation parameters ɵ. Based on projected parameters, the second portion generates a sample grid [14]. This portion generates the sampling grid, which is then delivered as input to the learnable interpolation algorithm in the third section, which produces the altered feature map O as an output. Part by part, we'll go through all you need to know in this area.

- **Localization network**:

For example, an input feature map with dimensions such as height and width are fed into the localization network, which creates output parameters such as spatial transformation. The network of localization will locate and localize N letters, words, or lines of text. It will be necessary to use an affine transformation matrix in order to apply rotations, translations, skew, and zoom to the input feature map in order to achieve oriented text detection. When it comes to text rotation, translation, and zoom, this system has a lot to learn.

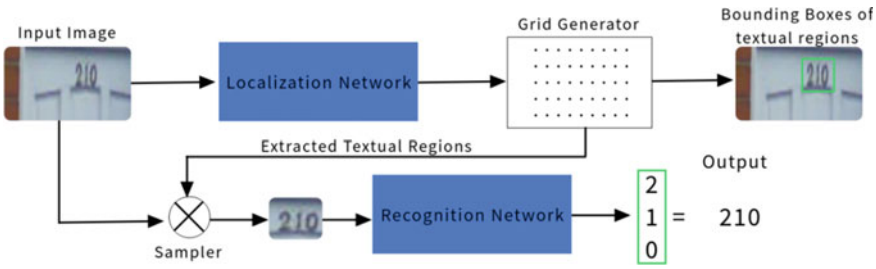

Fig. 1 Text detection and recognition using STN-OCR

STN-OCR uses a feed-forward CNN and an RNN to generate N affine transformation matrices. This network of localization makes use of the CNN model ResNet-50. The system's performance is superior to that of other network architectures, such as VGGNet, while using this network structure. As a result, it overcomes the problem of vanishing gradient and maintains a higher level of accuracy than alternative network structures. For the experiments, Batch Normalization was employed, and subsequently, RNN was used for the rest of the study. In this case, the RNN is a Bi-directional LSTM RNN. Hidden states are used to predict affine transformation matrices. BLSTM is primarily responsible for generating concealed states.

- **Localization Network Configuration**

ResNet architecture, or residual neural network, is utilized in the localization network. Pictures from this study will be sent to the network, which will then use them to locate the corresponding texts. The first layer of the network will use 32 filters to do a 3×3 convolution, the second layer will use 48 filters to accomplish the same convolution, and the third layer will use 48 filters to perform the same convolution. This process is followed by Batch Normalization and averaging 2×2 and stride two for each convolution layer. In each layer, ReLU is employed as an activation function. Batch Normalization is applied after each layer, followed by the usage of two residual layers with 3×3 convolution. Finally, a BLSTM with 256 neurons was applied to the last residual layer. A sampling grid with bounding boxes (BBoxes) retrieved for textual portions is constructed after the aforesaid model. Only the textual portion of the document, as seen in Fig. 2, is used to generate BBoxes.

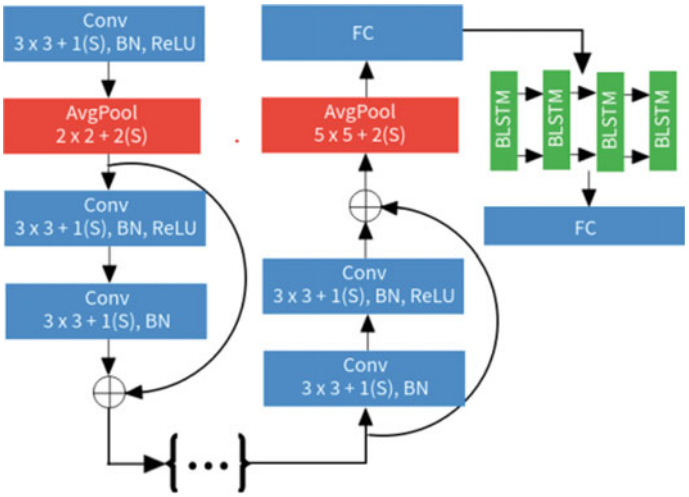

Fig. 2 Localization network

- ***GRID Generation***

Using the feature map as input, the system creates n grids of the input feature map I using grid G0 and the coordinates xw0 and yh0. This stage generates a total of N output grids, including the BBoxes of the network-located textual parts.

- ***Image sampling***

The values of feature map I were sampled at their corresponding coordinates on each of the N grids after N sampling grids were created with the grid generator in the second section. I can't fit these spots into the feature map grid since they don't make sense. As a result, bi-linear sampling was employed to select individuals from the regions that are closest to the centre of the population. The grid generator and picture sampler in action can be seen in Fig. 2. To choose picture pixels at a certain point in an image, the image sampler uses the grids generated by the grid generator. The vertices of the sampling grids are used to generate BBoxes automatically by this technique. Hence, the Spatial Transformer is formed by merging these three components: localization network, grid generation, and picture sampling, and can be employed in any region of a Deep Neural Network. This system begins with Spatial Transformer.

B. ***Text Recognition***

It returns N textual areas retrieved from the input image as a result of the text detection stage This stage of text recognition treats each of the N regions separately. CNN handles the processing of N regions. Because ResNet has been shown to produce better outcomes in text recognition systems, a ResNet variant is also used in this CNN. Text recognition is required to produce strong gradients for text detection. Probability distributions over label space are estimated at this stage. Probability distributions can be predicted using Softmax classifiers.

$$X^n = O^n \tag{1}$$

$$y_t^n = \text{soft max}(f_{\text{rec}}(x^n)) \tag{2}$$

For example, we get the output $f_{\text{rec}}(x)$ after convolution feature extraction.

Its configuration is identical to that of a localization network except for convolution filters. There are three convolutional layers totaling 32, 64, and 128 filters in this network.

C. ***Training Network***

An image training set X and a text file for each individual image are used to train the network/model in ICDAR 2015. Coordinates for the top-left, top-right, bottom-right, and bottom-left coordinates of × 1 and y1, × 2 and y2, × 3, y3, × 4, and Y4 in each image are included in each file. After learning localization and detecting possible text possibilities in the first step, the model employs labeling to identify the specific piece of text.

By calculating the loss of predicted text labels, error gradients are used to search and locate text regions. Some pre-training activities are required because we discovered that the model fails to merge multi-line texts into a picture. Optimizing the network during model training has a substantial influence. Adam optimizer is used after pre-training the network using Stochastic Gradient Descent (SGD) in order to improve the network's performance on more difficult tasks. The learning rate in the first step of text detection is kept constant for a longer period of time. This is leading to an improved ability to locate and identify textual sections. As a result, SGD is employed, and it performs admirably in this context. The next stage of text recognition involves learning to recognize text sections that have already been predicted in a prior stage.

4 Results and Discussions

What can be accomplished by utilizing this study structure is examined in this section. This investigation uses the ICDAR 2015 and SVHN benchmark datasets. A discussion of the aforementioned datasets follows.

- **ICDAR 2015 Dataset**: Robust Reading Competition makes use of the ICDAR 2015 dataset, which includes 1500 images and over 10,000 annotations. A total of 1000 photographs are utilized in the training process, and a further 500 images are used in the testing process. In addition, photographs can be annotated with text. There are three primary functionalities of ICDAR 2015: text recognition and word localization. Each image has a set of text-bound boxes (BBoxes) for localization. Each image's BBoxes is kept in its own file, separated just by a single line. To aid in automatic word recognition, BBoxes are provided in addition to the word itself.
- **SVHN Dataset**: Low-resolution photos and little data processing and formatting make the Street View House Numbers (SVHN) dataset a good benchmark. It could be compared to the MNIST database. Because it was compiled from house numbers in Google Street View pictures, this collection comprises a wide range of photographs, including blurry, low-resolution images. It is available in two formats: a chopped digits image format similar to MNIST and a complete home door image file with digit bounding boxes. Too many photos in this dataset: 73,257 for training, and 26,032 for testing. It's a mess.

Experiments on Datasets: ICDAR 2015 was the first dataset to be experimented with. For this dataset, the most difficult part is the variety of photos, which include a variety of background noises and clutter, as well as fuzzy photographs and low-resolution images. Pictures of the outcomes can be seen in Fig. 3.

With a Recall, Precision, and H-mean score of 64.2, 79.53, and 72.86%, the STN-OCR approach exceeds all other methods. The comparison is shown side by side in Fig. 4.

Fig. 3 Findings from the incidental scene text category of the 2015 international conference on documentary arts research

Fig. 4 STN-OCR performance

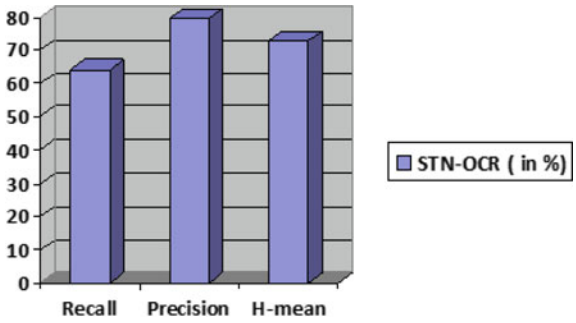

Table 1 STN-OCR performance

Method	Recall (%)	Precision (%)	H-mean (%)
STN-OCR	64.2	79.53	72.86
Baidu VIS	62.12	70.28	65.89
HoText_v1	62.26	67.25	66.78
FOTS	52.19	74.59	64.42

Table 2 Results on SVHN dataset

Method	Accuracy (%)
MaxoutCNN	95
ST-CNN	95.3
STN-OCR	97.8

According to the data in Table 1, the proposed strategy has produced superior outcomes when compared to the alternatives. On the SVHN dataset, the network design is evaluated to demonstrate that this model can be used for real data. House numbers in SVHN also contain noise. Finding, locating, and recognizing SVHN house numbers on a sampling grid was found to be a successful method of testing this study's network architecture. While initializing random weights were used to train the research model, weights from an already-trained network were used to initialize the localization network for optimum results. Better outcomes are generally obtained when using the localization network stage. A comparison of text recognition performance using real house numbers and the SVHN dataset is shown in Table 2.

This system's accuracy on SVHN improved after ICDAR 2015, when it reached 97.8%. Even if previous research has dealt with some of these conclusions, this study model works well with those photos. An image with a colour backdrop is processed in 2–3 s using Google's K80 GPU and 12 GB of RAM for testing purposes.

5 Conclusion

In this study, a single DNN was utilized to perform text detection and recognition (STN-OCR) utilizing recent benchmark datasets, such as ICDAR 2015. Two of the most important parts of this system are its text detection and recognition components. Text detection models are supplied into a text recognition network, which uses that network's output to recognize text areas in images. As a result, we were able to better detect text from several perspectives. According to the findings, our model outperforms current best practices by a wide margin on SVHN and ICDAR 2015 tests. Only whole sentences and lines are possible with this model. In the future, this model will be applied to other regional or well-known languages (such as Urdu/Hindi) and the geometric design will be adjusted to detect directly curved texts.

References

1. A. Alshanqiti, A. Bajnaid, A. Rehman, S. Aljasir, A. Alsughayyir, S. Albouq, Intelligent parallel mixed method approach for characterising viral Youtube videos in Saudi Arabia. Int. J. Adv. Comput. Sci. Appl. (2020)
2. Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, X. Bai, Textfield: Learning a deep direction field for irregular scene text detection. IEEE Trans. Image Process. (2019)

3. S. Khan, D.-H. Lee, M.A. Khan, A.R. Gilal, G. Mujtaba, Efficient edge-based image interpolation method using neighboring slope information. IEEE Access **7**, 133539–133548 (2019)
4. S.L. Xue, 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
5. A.R. Gilal, J. Jaafar, L.F. Capretz, M. Omar, S. Basri, I.A. Aziz, Finding an effective classification technique to develop a software team composition model. J. Softw. Evol. Process **30**(1), 1–12 (2018)
6. A. Sain, A.K. Bhunia, P.P. Roy, U. Pal, Multi-oriented text detection and verification in video frames and scene images. Neurocomputing **275**, 1531–1549 (2018)
7. C. Bartz, H. Yang, C. Meinel, See: Towards semi-supervised endto-end scene text recognition, in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (2018)
8. M. Liao, B. Shi, X. Bai, X. Wang, W. Liu, Textboxes: A fast text detector with a single deep neural network, in *Thirty-First AAAI Conference on Artificial Intelligence* (2017)
9. Y. Wei, Z. Zhang, W. Shen, D. Zeng, M. Fang, S. Zhou, Text detection in scene images based on exhaustive segmentation. Sig. Process. Image Commun. **50**, 1–8 (2017)
10. Y. Zhu, C. Yao, X. Bai, Scene text detection and recognition: Recent advances and future trends. Front. Comp. Sci. **10**(1), 19–36 (2016)
11. B. Xiong, K. Grauman. Text detection in stores using a repetition prior, in *Proceeding of the WACV* (2016)
12. S. Qin, R. Manduchi, A fast and robust text spotter, in *Proceeding of the WACV* (2016)
13. Z. Zhang, W. Shen, C. Yao, X. Bai, Symmetry-based text line detection in natural scenes, in *Proceeding of CVPR* (2015)
14. I. Posner, P. Corke, P. Newman, Using text-spotting to query the world, in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2010), pp. 3181–3186