IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEEP LEARNING FOR SCENE TEXT DETECTION: A SURVEY

[1] KHANAGHAVALLE.G.R, [2]KEERTHANA.S

[1] Department of Computer Science and Engineering,

[1]Sri Venkateswara College Of Engineering, Sriperumbedur, Tamil Nadu.

[2]Department of Computer Science and Engineering,

[2]CEG, Anna University, Chennai, Tamil Nadu.

Abstract: Text always plays an important role in human's life. Text in Natural Scene image is packed with rich symbolic information. Text detection aims at detecting the presence of text from the given input image. With the evolution of smart phones and the internet, a huge amount of image data is available for analysis. In Image Processing, Text Detection has become the most prominent research topic. This can be used in many applications such as Natural Scene Image understanding, developing applications for visually impaired people to read the text etc. This survey pays more attention to deep learning techniques which are used in recent years for text detection. In summary, this survey can serve as a good source of reference for researchers in the field of Text Detection.

Index Terms **- Deep Learning, Image processing, Text Detection.**

1. INTRODUCTION

Scene Text means the text which appears on the image captured by the camera in outdoor places such as road signs, license plate number, product packages, etc. Scene Text Detection is the most challenging task in Scene Text Recognition system. They are difficult because of several reasons such as complex background, small font sizes, and arbitrary orientations. Traditionally Text Detection is carried out using OCR technology for scanned document images. But the technique is not suitable for complex background images. With the emergence of Deep Learning many new techniques were developed to solve the problem.

Fig 1. Illustration of Scene Text Detection

Deep Learning is a branch of Machine Learning which has gained popularity in recent years. Deep Neural Networks forms the base of Deep Learning. A Deep Neural Network consists of an input layer, multiple hidden layers and an output layer. It converts the original data into high-level data by combining the lower level features learnt. Generally, Deep Learning uses a large amount of data to learn so it is able to achieve good results.

2. BACKGROUND

To understand the value of Scene Text Detection, it is important to provide background information about different methods in text detection and its challenges.

Classification of Text Detection Methods:

Text detection methods can be broadly into two types. They are primitive detection target and shape of the bounding target**.**

Classification Based on Primitive Detection Target: These methods can be again classified into three methods. They are the Character-Based method, Word-Based method and Text-line Based method. Character-Based methods try to detect the characters and merge them to form words. The word-based method detects the entire word. Text-line Based method uses symmetric characteristics of the text to detect the text-line.

Classification Based on Shape of the Bounding Target: This method detects the text by forming a bounding box around them. Horizontal bounding boxes method is more predominantly used to detect the words. Later Multi-Oriented bounding boxes are used to detect the text.

Challenges:

The complexity of text detection are the following:

- ✓ Characters in document images have a uniform size but characters in scene images have different size.
- ✓ Scene images always have a complex background which makes text detection a complex task.
- ✓ Various inference factors such as noise, blur and distortion occurs.
- ✓ Different font styles are used in the different outdoor environment.
- ✓ It becomes difficult to distinguish textual and non-textual images.
- ✓ Images captured with uneven lighting becomes difficult to detect text.

3. Recent Advances in Scene Text Detection

3.1. TextBoxes++

TextBoxes++ [4] is a fully trainable convolutional neural network to detect the text. It has only Convolution and Pooling layers. It can find arbitrary-oriented text via a novel regression method. It is based on the idea of the Object Detection algorithm. It detects the text in 6 stages. The network is adapted to predict the regression from the default boxes. Then densify boxes are used to cover the dense text region. It inherits the architecture of VGG-16. Multiple output layers are called as Text-Box layers. This network adopts an arbitrary sized image for both the training and testing phase. It uses 1*5 convolution filters instead of 3*3 filters.

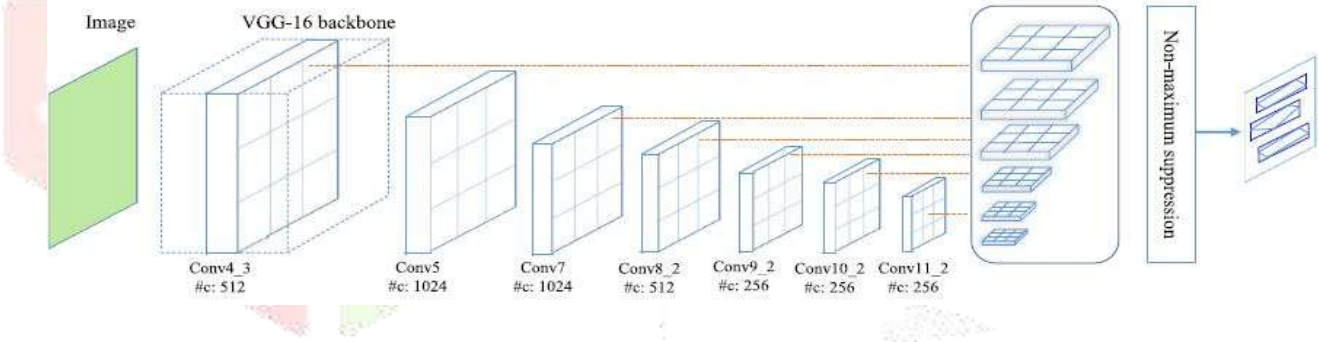

Fig 2. The architecture of TextBoxes++ [4]

3.2. Superpixel Based Stroke Feature Transform(SSFT)

Text region generally contains multiple characters. Those characters are extracted to form Candidate Text regions (CTR). Traditionally candidate character regions were extracted by finding the edges in the image. This approach was time-consuming. To overcome this problem superpixel Based stroke Feature Transform (SSFT) [8] was proposed. In this method first, the input image is resized and then smoothed with an edge filter. Then using the Simple Linear Iterative Clustering (SLIC) algorithm the image is segmented into K superpixels. These superpixels are clustered to form uniform superpixels. Euclidean Distance calculated between them is used to cluster them. Fast Edge Detection (FED) algorithm is used to detect the edges. Once the edges are detected, the background region is removed. Fig 3. shows the architecture of SSFT which has a low computational cost and greater robustness to noise. Then Deep Learning-Based Region Classification (DLRC) is used to remove the background region which still remains in the candidate regions.

3.3. Rotation Based Text Detection

Rotation Based Text Detection [2] has a Rotation Proposal Networks (RPN) and Rotation Region of Interest (RRoI) pooling layers. The framework has a VGG-16 Convolutional Neural Network in the front which is shared by two sibling branches i.e., RRPN and convolutional layer. RRPN has a classification layer and a regression layer. The feature maps are slided over to generate horizontal region proposals. These features are fed into the sliding branches. the ground truth of a text region is represented as rotated bounding boxes which are used for the

training phase. The RRoI max pooling layer converts the text proposals into a feature map. A classifier formed by two fully connected layers then classifies the features as background or text.

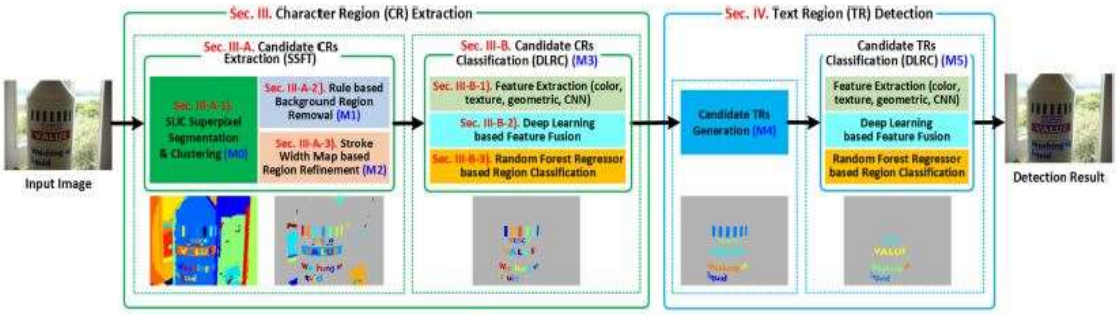

Fig 3. The architecture of SSFT

3.4. Multi-Oriented Scene Text Detection with Direct Regression

The proposed method [6] has four modules. they are feature extraction, multi-level feature fusion, multi-task learning and post processing. Convolution neural network is designed in such a way that the maximum receptive field is larger than the input image size. Because of this, regression can find a longer text.

Feature extraction method uses VGG-16, ResNet-50 and S-VGG. Multi-level feature fusion has a top-down approach. this provides feature maps of finer-resolution. It brings high efficiency which is accurate for boundaries localization. The feature maps are 1/4 of the input image. Multi-task output module produces a 4*4 region of an input image. After getting the output from multi-task learning, the non-textual regions are removed by the use of a classification task. To achieve efficient quadrilateral boundary, Direct Regression is used.

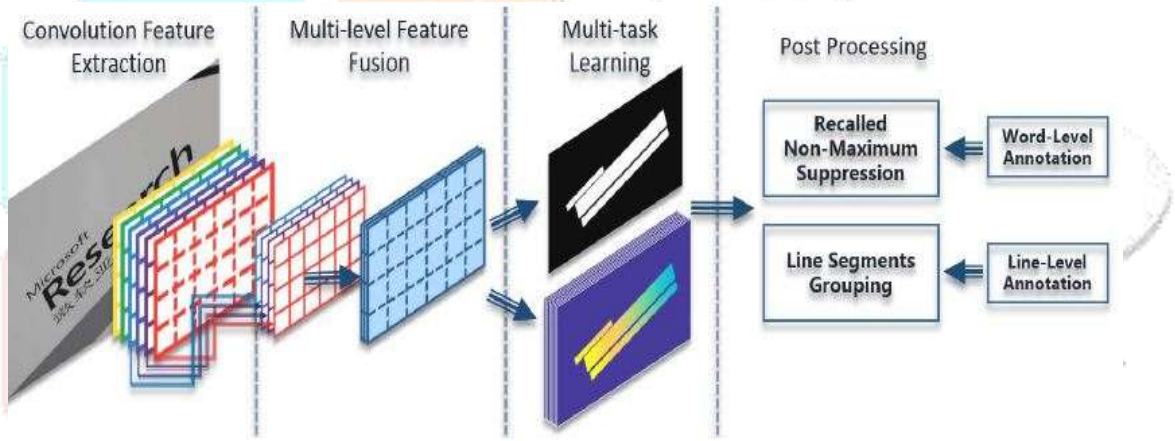

Fig 4. The pipeline of Multi-Oriented Scene Text Detection [6]

3.5. Feature Pyramid Based Text Proposal Network

Feature Pyramid Network [1] is an end-to-end trainable framework designed to detect the single character individually. It combines both Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The CNN part is used to extract the features from the input image and the RNN is used to encode the contextual information.

The FPN uses both top-down pathway and bottom-up pathway to learn features. This gives high-resolution and strong semantic features. RNN is used to reduce the false and missing inspections by making use of contextual information. Bi-LSTM encoding is used to consider both the context before and after the text. The internal state of each Bi-LSTM levels is connected to the FC layer to produce output. ROI pooling layer is used to classify the textual and non-textual region.

3.6. A Fast Text Detector using Knowledge Distillation

An end-to-end trainable model [5] is proposed to locate the multi-oriented scene text. It has two networks a student network and a teacher network. It inherits the complex VGGNet and PVANet. Knowledge distillation provides anew way of training the neural network. The knowledge of the teacher network is transferred to the student network with artificial annotations. For input images, the soft target generated by teacher network is more than the student network.

The teacher network is pretrained with parameters. The feature representation of the student network is the same as that of the teacher network. The teacher network is implemented using VGG-16 architecture and student network is implemented using PVANet architecture. The teacher network produces the convolution features for the input images. This text detector provides better accuracy for both horizontal and oriented text location.

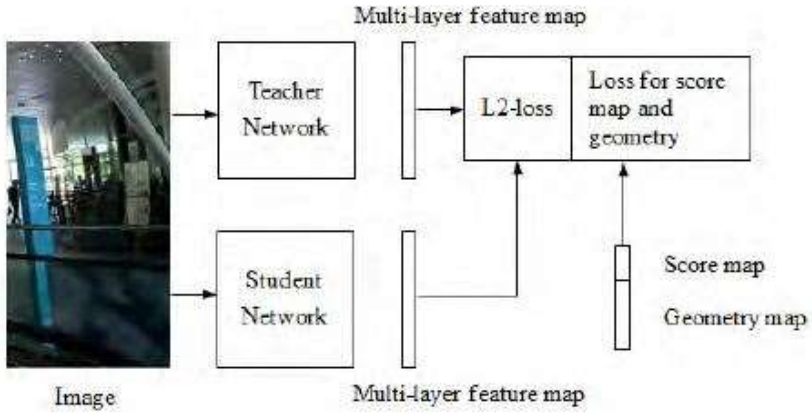

Fig 5. The architecture of Fast Text Detector[5]

3.7. Gradient-Inductive Segmentation Network with Contextual Attention

GISCA [3] is proposed to detect multi-oriented texts. It is an end-to-end fully trainable convolutional neural network. The is divided into four modules namely, Gradient Inductive module (GIM), Contextual Attention Module (CAM), U-Net network, and Pixel-level prediction module. This architecture inherits VGG-16 networks. U-shape feature fusion network is used for image segmentation. CAM aims to suppress the irrelevant background. CAM is proposed to boost the silent text areas which reduces the false positive. GIM updates the gradient information. It is designed to decouple the linear and non-linear expressions as two branches. The pixel link has two branches. One corresponds to text/ non-textual predictions and the other is used for age predictions. Finally, to detect text areas Connected components (CC) algorithm is employed.

3.8. SSD and Encoder-Decoder Network for Scene Text Detection

The framework [7] has two processes namely text detection and text localisation. Text detection detects the presence of text initially using SSD and text verification model eliminates the false localized text region using Encoder-Decoder network. The framework inherits the convolutional neural network with two parts. One part has 13-layers and the other part has 15-layer of convolution. Text detection layers use multiple feature maps to predict the bounding boxes. It combines multiple feature maps of different layers. Each position of the feature map is mapped to the original image. the number of default bounding boxes is increased from the traditional method. Natural Scene Images has a complex background because of which many text regions are misjudged. To overcome this problem Encoder-Decoder network is used. This network is composed of CNN and BiGRU network. Encoder network encodes the image into feature vector sequence. The decoder network converts the vector into words.

Table 1: shows the comparison of various techniques.

Algorithms	Dataset used	F-Score
Minghui Liao et al.[4]	ICDAR2015, Synth dataset, Coco-Text Dataset	0.829
Youbao Tang et al.[8]	ICDAR2011, ICDAR 2013, SVT Dataset	0.879
Jiangi Ma et al.[2]	MSRA-TD500, ICDAR 2015	0.80
Wenhao He et al.[6]	ICDAR 2013, ICDAR 2015, MSRA-TD500, MLT-17, CASIA-10K	0.91
Fagui Liu et al.[1]	ICDAR 2013, ICDAR 2015, USTB-1K	0.925
Peng Yang et al.[5]	ICDAR 2015, COCO-Text Dataset, ICDAR 2013	0.90
Meng Cao et al.[3]	ICDAR 2013, ICDAR 2015, MSRA-TD500	0.921
Xue Gao et al.[7]	ICDAR 2017, RCTW 2017	0.784

4. CONCLUSION

Scene Text Detection is the most complex area of research in Computer Vision. This paper presents an extensive survey of Scene Text detection frameworks available in Deep Learning. Although the research has more progress in recent years, it still has some problems to explore. We can still make improvements in various aspects. Future research work can be conducted in automatically annotating the images and in Visual Question and Answering System.

REFERENCES

[1] Fagui Liu, Cheng Chen, Dian Gu, Jingzhong Zheng.:FTPN: Scene Text Detection With Feature Pyramid Based Text Proposal Network. In:IEEE Access, Vol.7, pp. 44219 - 44228. IEEE(2019)

[2] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, Xiangyang Xue.:Arbitrary-Oriented Scene Text Detection via Rotation Proposals. In:IEEE Transactions on Multimedia , Vol. 20, pp.3111 - 3122. IEEE(2018)

[3] Meng Cao, Yuexian Zou, Dongming Yang, Chao Liu. In:GISCA: Gradient-Inductive Segmentation Network With Contextual Attention for Scene Text Detection. In:IEEE Access, Vol.7, pp. 62805 - 62816, In:IEEE Access(2019)

[4] Minghui Liao, Baoguang Shi, Xiang Bai.:TextBoxes++: A Single-Shot Oriented Scene Text Detector. In: IEEE Transactions on Image Processing, Vol. 27, pp. 3676 - 3690. IEEE(2018)

[5] Peng Yang, Fanlong Zhang, Guowei Yang.:A Fast Scene Text Detector Using Knowledge Distillation. In:IEEE Access, Vol.7, pp.22588 - 22598. IEEE(2019)

[6] Wenhao He, Xu-Yao Zhang, Fei Yin, Cheng-Lin Liu.:Multi-Oriented and Multi-Lingual Scene Text Detection With Direct Regression. In:IEEE Transactions on Image Processing , Vol.27, pp. 5406 - 5419. IEEE(2018)

[7] Xue Gao, Siyi Han, Cong Luo.:A Detection and Verification Model Based on SSD and Encoder-Decoder Network for Scene Text Detection. In:IEEE Access, Vol. 7, pp. 71299 - 71310, In: IEEE(2019)

[8] Youbao Tang, Xiangqian Wu.:Scene Text Detection Using Superpixel-Based Stroke Feature Transform and Deep Learning Based Region Classification. In:IEEE Transactions on Multimedia, Vol. 20, pp. 2276-2288. IEEE(2018)