**CRAFTY OCR:**

**A TEXT DETECTION EXPERIMENT**

**Abstract**

The goal of this paper is to experiment, demonstrate and bring up possible suggestion for the problem of effective text detection and recognition for distorted character images. Bearing the goal of effectiveness in mind, we looked to implement the state-of-the-art model in the discipline and implement it. While there are many end-to-end deep learning models which carry out the task of text detection and text recognition simultaneously, which improve robustness of the text detector. However, the domain of curved text images still proves to be a challenge for these 2-in-1 models. Therefore, we wish to dig deep into the detection part for best result, with a deep text recognition model added on later in the process. To illustrate our work in a clear and concise manner, the paper will be divided into four part. In the first part, we will give a quick scan of related work on the same problem. Following that, the dataset in use model architecture will be described in details so that readers can get a better grasp of our work. The third part will be devoted to explaining the text recognition framework deployed after text’s bounding boxes have been found. In the fourth and final part of our paper, we will give readers a brief summarization of the report, discuss the limitations of the models and the problems we have not tackled, as well as open new prospects for future work. For idea, mathematical concept and source code, we derive mostly from the work proposed in the paper “Character Region Awareness for Text Detection” of Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee from Clova AI, as well as the text recognition framework.

In the first part, Text Detection, we attempt replicate and improve a neural network-based text detection. Old text detection methods are based on based on pre-fixed bounding boxes have a lot of limitations in detecting real-life arbitrary texts. The new way proposed in the paper can effectively detect text area by exploring each character and affinity between characters. To overcome the lack of individual character level annotations, the proposed framework exploits both the given character-level annotations for synthetic images and the estimated character-level ground-truths for real images acquired by the learned interim model. In order to estimate affinity between characters, the network is trained with the newly proposed representation for affinity. In the original paper, extensive experiments have been carried out on six benchmarks, including the TotalText and CTW-1500 datasets which contain highly curved texts in natural images. However, due to the computational and time limitation, we were only able to focus our work on the TotalText language, which features English language. Future works based on this report can shift the attention to other languages of the same characteristics.

**I. Literature reviews**

A sub-field of computer vision, scene text detection is becoming a widespread application of computer vision field because of its numerous applications, such as instant translation, image retrieval, blind-navigation, to name but a few. These methods mainly train their networks to localize word-level bounding boxes. However, they have found success mostly on regular-shaped text images, in case the images are curved, deformed, arbitrary or extremely long, they bare give a satisfactory performance. Thus, the concept of character-level awareness has been proposed to tackle this problem. By linking the successive characters in a bottom-up manner, this approach provides better performance for out-of-the-ordinary text instances. However, it has seen little implementing success compared to older approaches, due the the lack of satisfactory datasets and the cost required to obtain such training materials.

In the paper “” paper proposed a novel text detector that localizes the individual character regions and links the detected characters to a text instance. Our framework, referred to as CRAFT for Character Region Awareness For Text detection, is designed with a convolutional neural network producing the character region score and affinity score. The region score is used to localize individual characters in the image, and the affinity score is used to group each character into one instance.

The major trend in scene text detection before the emergence of deep learning was bottom-up, where handcrafted features were mostly used – such as MSER or SWT– as a basic component. However, deep learning-based text detectors have been overtaking the scene, inheriting achievement form object detection/segmentation methods like SSD , Faster R-CNN, and FCN. Some of the most note-worthy methods include:

* **Regression-based text detectors** Various text detectors using box regression adapted from popular object detectors have been proposed. Unlike objects in general, texts are often presented in irregular shapes with various aspect ratios. To handle this problem, TextBoxes modified convolutional kernels and anchor boxes to effectively capture various text shapes. DMPNet tried to further reduce the problem by incorporating quadrilateral sliding windows. In recent, Rotation-Sensitive Regression Detector (RSDD) which makes full use of rotation-invariant features by actively rotating the convolutional filters was proposed. However, there is a structural limitation to capturing all possible shapes that exist in the wild when using this approach.
* **Segmentation-based text detectors** Another common approach is based on works dealing with segmentation, which aims to seek text regions at the pixel level. These approaches that detect texts by estimating word bounding areas, such as Multi-scale FCN, **Holistic-prediction**, and **PixelLink** have also been proposed using segmentation as their basis. SSTD tried to benefit from both the regression and segmentation approaches by using an attention mechanism to enhance text related area via reducing background interference on the feature level. Recently, TextSnake was proposed to detect text instances by predicting the text region and the center line together with geometry attributes.
* **End-to-end text detectors:** An end-to-end approach trains the detection and recognition modules simultaneously so as to enhance detection accuracy by leveraging the recognition result. FOTS and EAA concatenate popular detection and recognition methods, and train them in an end-to-end manner. Mask TextSpotter took advantage of their unified model to treat the recognition task as a semantic segmentation problem. It is obvious that training with the recognition module helps the text detector be more robust to text-like background clutters.

**II. Text detection framework**

**II.1. Methodology**

The objective of this CRAFT method is to precisely localize each individual character in natural images. To this end, a deep neural network was trained to predict character regions and the affinity between characters.

**II.2. Architecture**

A fully convolutional network architecture based on VGG-16 [34] with batch normalization is adopted as the backbone of the architecture. The model has skip connections in the decoding part, which is similar to U-net [31] in that it aggregates low-level features. The final output has two channels as score maps: the region score and the affinity score. The network architecture is illustrated in Figure 1.

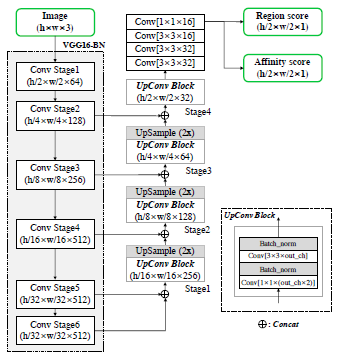


Figure 1: The model architecture of CRAFT framework

**II.3. Dataset description**

The dataset in use here is the Total Text Dataset. The basic description of the dataset is as follow: “Total Text Dataset - ICDAR 2017. It consists of 1555 images with more than 3 different text orientations: Horizontal, Multi-Oriented, and Curved, one of a kind.” To illustrate, the dataset is comprised of almost entirely irregular-shaped formations, suited for the question at hand, which is to detect arbitrarily-shaped text images, as opposed to older text dataset which consists mainly of regular text shapes.

On the plus side, the model is truly one of a kind, as stated by its brief description, every text image in the set has its own shape formation, which aims to reflect real-life text images found in the wild. This makes for a truly diverse, practical dataset. However, the downside is that the number of instances are few compared to other public datasets in the same domains. Moreover, the dataset does not focus on featuring text images under more unfavorable conditions, such as lack of lighting, obscurity, etc. Nonetheless, the quality of the data points in terms of shape diversity is good enough to carry out the work under discussion in the report.

**II.4. Code and Dependency Description**

To implement this model, we use Python 3.6. Some other dependencies such as Pytorch and CUDA are also required.

**III. Text recognition framework**

This part described the subsequent work following the detecting results. Basically, for each raw image, the output of text detector will be the text images with the bounding box, and the co-ordinates of that corresponding bounding box. We use the co-ordinates of the bounding box to crop the image then apply the text recognizer to detect the words therein.

The framework of the text recognizer is the STR framework, which allow for multiple modules to combine and contribute in terms of speed, accuracy and memory demand. The modules in use is TPS, Resnet, Bi-LSTM and Atttn.

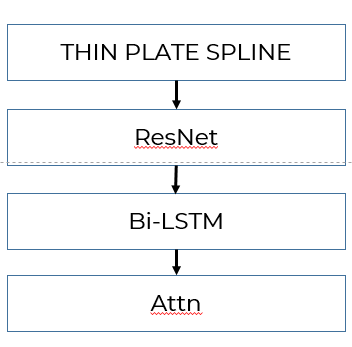


Figure 2: The architecture of the framework in use.

**IV. Summary of the work and recommendation.**

**IV.1 Result**

As expected, the detector performs satisfactorily, albeit with a few mistakes if the font type is too difficult to detect, or when there are some Chinese characters thrown in the mix.

As for the text recognizer, it performs well enough with single-word instances. As for the multiple-word instances, it still has much to be desired.

To demonstrate the result, we use the picture taken from some of the testing image dataset. The first column shows the images with the bounding box, the second shows the text recognition result and the third column show the word result.

|  |  |  |  |
| --- | --- | --- | --- |
| **Detected text (in bounding box)** | **Text recognition result (uncropped)** | **Text recognition result (cropped)** | **Correct text** |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img92.jpg | FISTER | Sucking  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img92.jpg | RESTORAN  MAKANAN  LAUT  Jeti  Seafood |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img95.jpg | the | VATESCOVE  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img95.jpg | UNITED  STATES  OF  AMERICA |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img98.jpg | From | a  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img98.jpg | CHELSEA  FOOTBALL |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img191.jpg | the | Subject  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img191.jpg | Pryde’s  Old  Westport  Kitchen  Home  Acsessories |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img192.jpg | the | Nonsinctranced  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img192.jpg | MORNING  CALL  COFFEE  STAND |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img298.jpg | and | From  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img298.jpg | GURU  NANAK  ENGINEEING  COLLEGE |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img299.jpg | From | HOTELRECEPTION  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img299.jpg | HOTEL  RECEPTION |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img393.jpg | a | PUZZLEMANSION  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img393.jpg | GUINNESS  WORLD  RECORDS  WORLD’S  LARGEST  JIGSAW  PUZZLE  PUZZLE  MANSION  BRGY  ASISAN  TAGAYTAY  CITY |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img492.jpg | FRESTES | ERECTED  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img492.jpg | ERECTED |
| E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\New folder\res_img494.jpg | CONTERSING | ISTANBUL  E:\BT_Cuoikhoa_RubikAI\deep-text-recognition-benchmark-master\demo_image\res_img494.jpg | ISTANBUL |

**V.2. Limitations and possible use case:**

As the text detector requires heavy setup, it may not be suitable for integrating into more lightweight devices. Some improvement work is required in order to see this detector be applied in real life. Also, more training data for better capability is also desirable.

As for the text recognizer, while it works well on single-word images, it still need enhancements to work on images with multiple images, or to recognize strings of word. As the model is still in its experimental stage, we can expect a great deal from this model in the future.

As for the possible use cases, the model would work well with the task of auto navigations. Namely, if the text detection and recognition is integrated into the vision of self-driving vehicles, it will be a great enhancement for the address or location recognition ability of the vehicles. Another use case is to recognize a location on an online based on its logo or address picture. This model will be very useful to support smart hardware systems.