htw.

**Hochschule für Technik
und Wirtschaft Berlin**

University of Applied Sciences

Text Detection in Natural Scene Images using CRAFT and EAST

Henry Febrian

Module: *Computergrafik und Bildverarbeitung*

Lecturer: Erik Rodner

September 30, 2021

Abstract

To be edited at a later date.

Contents

1 Introduction

Text is arguably one of the most essential form of communication. According to Long et al. (2020):

> *"As the written form of human languages, text makes it feasible to reliably and effectively spread or acquire information across time and space. In this sense, text constitutes the cornerstone of human civilization."* (Long et al. 2020)

In the modern world, text as a medium of communication is not only consumed by humans but has claimed its place in the world of technology. However, text detection in natural scene images is proven to be challenging. Compared to detecting text on handwritten materials, the randomness of a natural scene is a big hurdle to overcome. This paper aims to test and evaluate the two proposed methods on their performance in detecting natural-scene texts. It begins by observing the interference found on natural scenes followed by introducing and giving an overview of CRAFT and EAST. It then provides an explanation of the evaluation dataset. Subsequently, the performance review of the algorithm will be presented, which is obtained by testing the preferred method on the aforementioned dataset.

2 Challenges of Natural Text Detection

Natural scene images could be classified as images which are taken in uncontrolled environments, with any device ranging from smartphones to professional cameras. These images are snapshots of things in the real world.

(a) A scenery   (b) A billboard

Figure 1: Samples of natural scene images.

The random nature of the real world, combined with the diversity of available devices introduced some factors which make natural text detection a greater

challenge than detecting structured text in documents. Mancas-Thillou & Gosselin (2007) mentioned some conditions that are found in natural scene which may significiantly impact text detection procedure. They are:

- Raw sensor image and sensor noise

- Viewing angle

- Blur

- Lighting

- Resolution

- Non-paper objects

- Non-planar objects

Although devices have evolved to a point where most of handheld devices are capable of shooting in a high resolution, low-end handheld cameras and older models still struggle in this sector. Uncontrolled environment, combined with the possible lack of stabilization from the equipment can cause blur (Rosebrock 2018). Also, there are countless factors such as the time of the day, weather, camera flash, and many others which may impact the lighting conditions, further hindering the ability to detect text. Non-paper objects such as glass and plastic may reflect images, and non-planar objects such as text wrapped around a bottle becomes distorted and deformed (Rosebrock 2018). Additionally, there might be patterns that are extremely similar to text, or occlusions caused by foreign objects, which may potentially lead to confusion and mistakes (Long et al. 2020).

3 Methodology

3.1 Overview of EAST

According to Zhou et al. (2017), the key component of EAST is a neural network model, which is trained to directly predict the existence of text instances and their geometries from full images (Zhou et al. 2017). Hence, the abbreviation EAST: Efficient and Accurate Scene Text Detector.
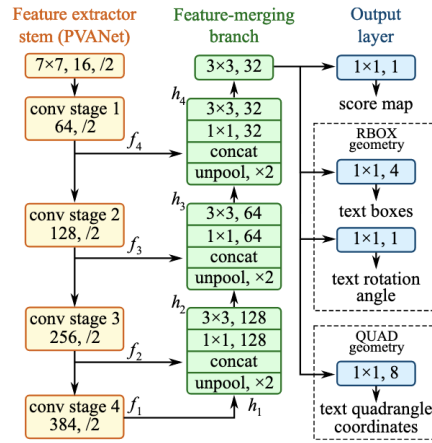

Figure 2: Schematic view of EAST, adopted from Figure 3 of Zhou et al. (2017)

As claimed by Zhou et al. (2017), EAST is among the most efficient text detectors that achieve state-of-the-art performance on benchmarks. The idea of the network is described shortly below:

> *"... we adopt the idea from U-shape (Ronneberger et al. 2015) to merge feature maps gradually, while keeping the upsampling branches small. Together we end up with a network that can both utilize different levels of features and keep a small computation cost."* (Zhou et al. 2017)

It is capable of predicting text on 720p images, running at an average of 13 FPS. The fastest setting, which reached a speed of 16.8 FPS was achieved on a combination of their algorithm with PVANET on 720p images using NVIDIA Titan X graphics card. (Zhou et al. 2017). The training of the network is further described as follow:

> *"The network is trained end-to-end using ADAM (Kingma & Ba 2014) optimizer. To speed up learning, we uniformly sample 512x512 crops from images to form a minibatch of size 24. Learning rate of ADAM starts from 1e-3, decays to one-tenth every 27300 minibatches, and stops at 1e-5. The network is trained until performance stops improving."* (Zhou et al. 2017)

5

3.2 Overview of CRAFT

CRAFT stands for Character Region Awareness for Text Detection. According to Baek et al. (2019), CRAFT is a novel text detector which localizes the individual character regions and links the detected characters to a text instance (Baek et al. 2019).
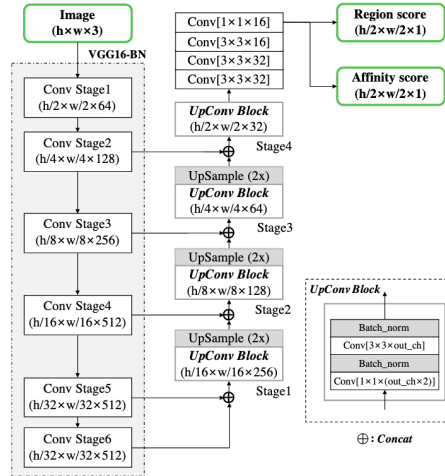

Figure 3: Structure of Craft, adopted from Baek et al. (2019)

As stated in the original paper, CRAFT adopts a fully convolutional network architecture based on VGG-16 (Simonyan & Zisserman 2015) with batch normalization as its backbone (Baek et al. 2019). It is further described as follow:

> *"Our model has skip connections in the decoding part, which is similar to U-Net (Ronneberger et al. 2015) 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 schematically illustrated in figure 3 (Baek et al. 2019)."*

4 Results and Evaluation

asdasdssada

5 Conclusion

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

6   Section about quotations

In this section, an example for a literal quotation is given.

> *"A persona is a rich picture of an imaginary person who represents your core user group."* (Baek et al. 2019)

Sometimes you might want to make use of the authors name within the text. Before, we used the command `citep{}`, which creates the brackets around author name and year. You can also use the `cite` command like this:

Baek et al. (2019) defined the concept of persona as follows:

> *"A persona is a rich picture of an imaginary person who represents your core user group."* (Zhou et al. 2017)

You may notice, that this increases the readability of the text.

According to APA format[1] there are some rules, when and how to include page numbers, when referring to literature.

> *"Include page numbers for any citations in the text of your paper that include direct quotations or refer to a specific part of the work you are referencing. Direct quotations must include a page number as part of the citation. The quoted material should be followed by a citation in parentheses that gives the author's name, the year in which the work was published, and the page number from which the quoted material appears."* (Long et al. 2020)

Check out the example and recommendations of Long et al. (2020) on `http://www.ehow.com/how_5689799_cite-numbers-apa-format.html`. In LATEXyou can include the pages very easy. For example:

stated:

> *"We hope that our preliminary attempts to begin answering the question will convince the reader, not necessarily that our views are correct, but that the question was and is well worth asking"* (Baek et al. 2019)

Note that in the first reference, we used `citet[]{}` in order to have brackets just around year and page number; later we used `citep[]{}`.

7   Section about references within the document

If you want to refer to you own chapters, figures, tables or the like, you can make use of the `ref{}` command, for example:

- section 6 on page 7

---

[1] American Psychological Association (APA)

7.1   Subsection within Foundations

Lorem ipsum dolor sit amet, consetetur Mancas-Thillou & Gosselin (2007) sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

7.2   Another subsection within Foundations

Lorem ipsum dolor sit amet, consetetur sadipscing Rosebrock (2018) elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

References

Baek et al. 2019
> BAEK, Youngmin; LEE, Bado; HAN, Dongyoon; YUN, Sangdoo; LEE, Hwalsuk: *Character Region Awareness for Text Detection*. `https://arxiv.org/pdf/1904.01941.pdf`. Version: 2019. – Last accessed 14 September 2021

Kingma & Ba 2014
> KINGMA, Diederik P.; BA, Jimmy: *Adam: A Method for Stochastic Optimization*. `https://arxiv.org/pdf/1412.6980.pdf`. Version: 2014

Long et al. 2020
> LONG, Shangbang; HE, Xin; YAO, Cong: *Scene Text Detection and Recognition: The Deep Learning Era*. `https://arxiv.org/pdf/1811.04256.pdf`. Version: 2020. – Last accessed 7 September 2021

Mancas-Thillou & Gosselin 2007
> MANCAS-THILLOU, Céline; GOSSELIN, Bernard: *Natural Scene Text Understanding*. `https://pdfs.semanticscholar.org/c6de/0b0485e4ebc6df300669969ed743b86b511e.pdf?_ga=2.82453078.1125900540.1632067547-1687845596.1632067547`. Version: 2007. – Last accessed 20 September 2021

Ronneberger et al. 2015
> RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015), pages 234–241

Rosebrock 2018
> ROSEBROCK, Adrian: *OpenCV Text Detection (EAST Text Detector)*. `https://www.pyimagesearch.com/2018/08/20/opencv-text-detection-east-text-detector/`. Version: 2018. – Last updated July 7, 2021; Last accessed 21 September 2021

Simonyan & Zisserman 2015
> SIMONYAN, Karen; ZISSERMAN, Andrew: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: *ICSR* (2015)

Zhou et al. 2017
> ZHOU, Xinyu; YAO, Cong; WEN, He; WANG, Yuzhi; ZHOU, Shuchang; HE, Weiran; LIANG, Jiajun: *EAST: An Efficient and Accurate Scene Text Detector*. `https://arxiv.org/pdf/1704.03155.pdf`. Version: 2017. – Last accessed 14 September 2021