Electronic Letters on Computer Vision and Image Analysis 0(0):1-7, 2019

Fully Convolutional Networks for Text Understanding in Scene Images

* Dena Bazazian

*Advisors: * Dimosthenis Karatzas and + Andrew D. Bagdanov*
** Centre de Visió per Computador (CVC), Universitat Autònoma de Barcelona, Spain*
+ Media Integration and Communication Center (MICC), University of Florence, Italy
Permanent link of thesis: https://www.educacion.gob.es/teseo/mostrarRef.do?ref=1717635

Received 25 September 2019; revised 10 October 2019; accepted 1 November 2019

Abstract

Text understanding in scene images has gained plenty of attention in the computer vision community and it is an important task in many applications as text carries semantically rich information about scene content and context. For instance, reading text in a scene can be applied to autonomous driving, scene understanding or assisting visually impaired people. The general aim of scene text understanding is to localize and recognize text in scene images. Text regions are first localized in the original image by a trained detector model and afterwards fed into a recognition module. The tasks of localization and recognition are highly correlated since an inaccurate localization can affect the recognition task. The main purpose of this thesis is to devise efficient methods for scene text understanding. We investigate how the latest results on deep learning can advance text understanding pipelines. Recently, Fully Convolutional Networks (FCNs) and derived methods have achieved a significant performance on semantic segmentation and pixel level classification tasks. Therefore, we took benefit of the strengths of FCN approaches in order to detect and recognize text in natural scenes images.

Key Words: Text Understanding, Text Detection, Word Spotting, Fully Convolutional Network (FCN), Scene Images.

1 Introduction

The main purpose of this thesis is to devise efficient methods for scene text understanding. We investigate how the latest results on deep learning can advance text understanding pipelines. Recently, Fully Convolutional Networks (FCNs) [7] and derived methods have achieved a significant performance on semantic segmentation and pixel level classification tasks. Therefore, we took advantage of the strengths of FCN approaches in order to detect text in natural scenes. In this thesis we have focused on two challenging tasks of scene text understanding which are Text Detection and Word Spotting. For the task of text detection, we have proposed an efficient text proposal technique in scene images. We have considered the Text Proposals method [1] as the

Correspondence to: <dena.bazazian@cvc.uab.es>

Recommended for acceptance by <name>
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

baseline which is an approach to reduce the search space of possible text regions in an image. In order to improve the Text Proposals method we combined it with Fully Convolutional Networks to efficiently reduce the number of proposals while maintaining the same level of accuracy and thus gaining a significant speed up. Our experiments demonstrate that this text proposal approach yields significantly higher recall rates than the baseline text localization techniques, while also producing better quality localization [2, 3]. We have also applied this technique on compressed images from wearable egocentric cameras [6]. For the task of word spotting, we have introduced a novel mid-level word representation method. We have proposed a technique to create and exploit an intermediate representation of images based on text attributes which roughly correspond to character probability maps. Our representation extends the concept of Pyramidal Histogram Of Characters (PHOC) [8] by exploiting Fully Convolutional Networks to derive a pixel-wise mapping of the character distribution within candidate word regions. We call this representation the Soft-PHOC [4]. Furthermore, we show how to use Soft-PHOC descriptors for word spotting tasks through an efficient text line proposal algorithm. To evaluate the detected text, we propose a novel line based evaluation along with the classic bounding box based approach. We test our method on incidental scene text images which comprises real-life scenarios such as urban scenes. The importance of incidental scene text images is due to the complexity of backgrounds, perspective, variety of script and language, short text and little linguistic context. All of these factors together makes the incidental scene text images challenging.

2 Improving Text Proposals by FCN

Class-specific text proposal algorithms can efficiently reduce the search space for possible text object locations in an image. In this work we combine the TextProposals algorithm [1] with Fully Convolutional Networks to efficiently reduce the number of proposals while maintaining the same high level of recall. Hence, this technique leads to gain a notable speed up [2, 3]. Experiments demonstrate that our text proposal approach yields significantly higher recall rates than state-of-the-art text localization techniques, while also producing better quality localization. Our results on the ICDAR 2015 Robust Reading Competition (Challenge 4) and the COCO-text datasets show that, when our technique combined with strong word classifiers, this recall margin leads to state-of-the-art results in end-to-end scene text recognition. A diagram of the architecture of our proposed framework is shown in Figure 1.

3 Word Spotting

Word spotting in natural scene images has many applications in scene understanding and visual assistance. To this end, in this thesis we address the problem of unconstrained Word Spotting in scene images. We have proposed two different techniques for word spotting, we call the first technique “Strong Character Labeling” [5] and the second one “Soft-PHOC” [4]. The comparison of these two techniques is shown in Figure 2.

In the first technique “Strong Character Labeling”, we trained an FCN-based model with a strong character level labeling to produce heatmaps of all the character classes. Afterwards, we employ the Text Proposals approach; then, via a bounding box classifier we detect the most likely bounding box for each query word based on the character attribute maps.

In the second technique “Soft-PHOC”, we propose a pipeline to create and exploit an intermediate representation of images based on text attributes which are character probability maps. Our representation extends the concept of the Pyramidal Histogram Of Characters (PHOC) by exploiting Fully Convolutional Networks to derive a pixel-wise mapping of the character distribution within candidate word regions. We call this representation the Soft-PHOC. A description of Soft-PHOC annotation by an example is shown in Figure 3. Furthermore, we show how to employ Soft-PHOC descriptors for word spotting tasks in egocentric camera streams through an efficient text line proposal algorithm. This is based on the Hough Transform over character attribute maps

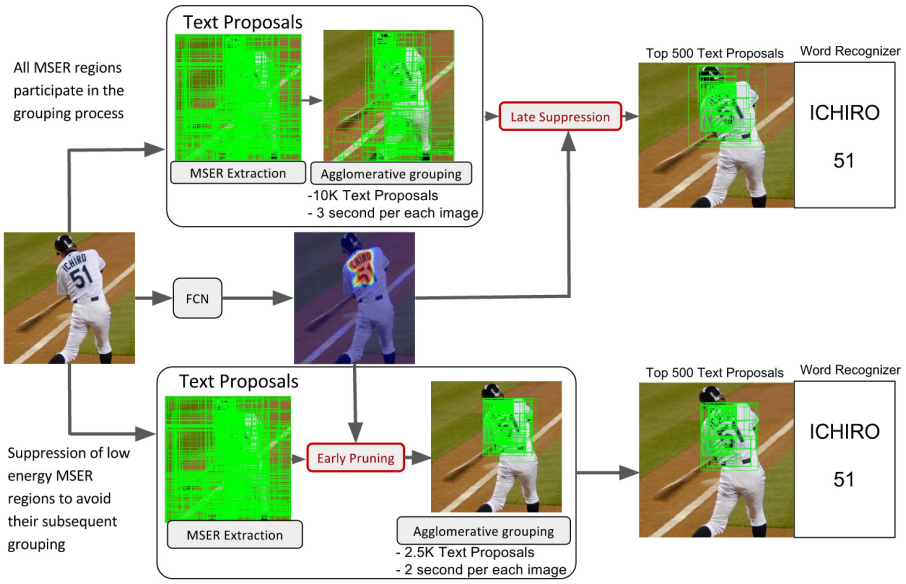

Figure 1: Comparison of the late suppression and early pruning strategies. The Late Suppression strategy (top) allows the text proposals algorithm to generate all hypotheses before filtering, while the Early Pruning strategy is tightly integrated with the Text Proposals algorithm guiding it to only generate relevant hypotheses.

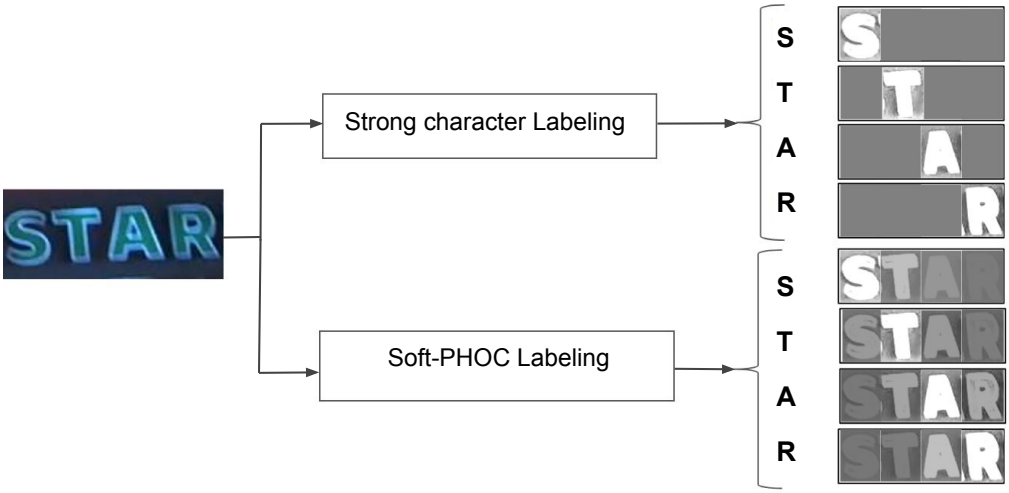

Figure 2: Comparison between the Strong Character Labeling and Soft-PHOC labeling strategies. Strong Character Labeling considers each character individually, regardless of the other characters in each word. Soft-PHOC labeling considers instead a soft probability distribution of all the characters in each word.

followed by scoring using Dynamic Time Warping (DTW). We evaluate our results on ICDAR 2015 Challenge 4 dataset of incidental scene text captured by an egocentric camera.

4 Conclusion

In this PhD dissertation we have addressed the problem of text understanding in scene images at different levels, from “where is the text?” to “what is written there?”, from the level of localizing the text to the level of recognizing the text in scene images. To achieve text understanding we developed robust Fully Convolutional Networks (FCNs) that included tools for both the text detection and word spotting tasks in scene images.

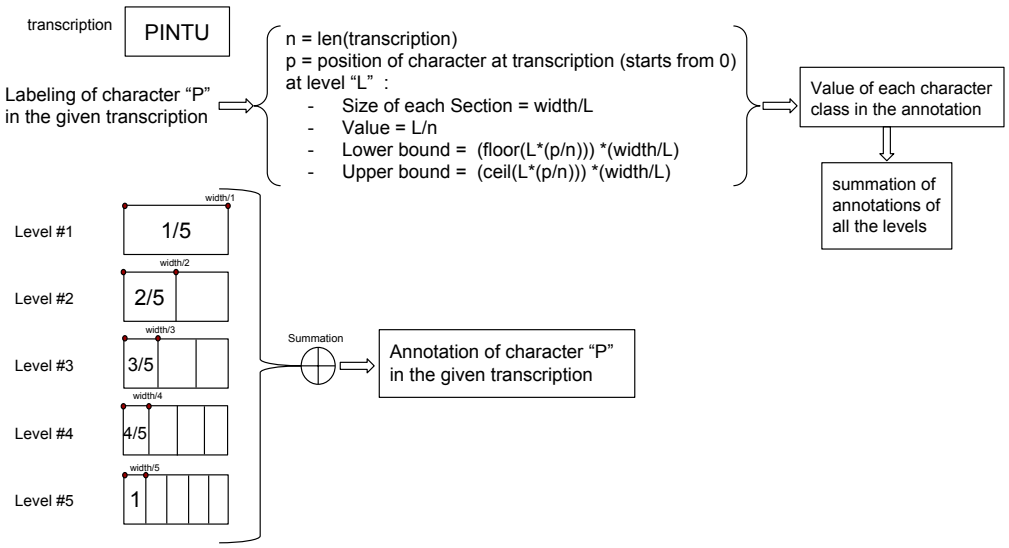

Figure 3: Soft-PHOC annotation. For instance, if the transcription is "PINTU", we show how we can define the annotation of class "P" for the given transcription based on the value at each level of Soft-PHOC descriptor.

References

[1] L. Gomez and D. Karatzas. TextProposals: a text-specific selective search algorithm for word spotting in the wild.Pattern Recognition, pages 60–74, 2017.

[2] D. Bazazian, R. Gomez, A. Nicolaou, L. Gomez, D. Karatzas, and A. Bagdanov. Improving text proposals for scene images with fully convolutional networks. In Proc. International Conference on Pattern Recognition, (DLPR workshop), arxiv:1702.05089, 2016.

[3] D. Bazazian, R. Gomez, A. Nicolaou, L. Gomez, D. Karatzas, and A. Bagdanov. Fast: Facilitated and Accurate Scene Text proposals through FCN guided pruning. In Pattern Recognition Letters, 2017.

[4] D. Bazazian, D. Karatzas, and A. Bagdanov. Soft-PHOC descriptor for end-to-end word spotting in egocentric scene images. In European Conference on Computer Vision (EPIC workshop), 2018.

[5] D. Bazazian, D. Karatzas, and A. Bagdanov. Word spotting in scene images based on character recognition. In Proc. Computer Vision and Pattern Recognition Workshop, pages 1872–1874, 2018.

[6] L. Galteri, D. Bazazian, L. Seidenari, A. Bagdanov, M. Bertini, A. Nicolaou, D. Karatzas, and A. Bimbo. Reading text in the wild from compressed images. In Proc. International Conference on Computer Vision Workshops, pages 2399–2407, 2017.

[7] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence.39 (4), pages 640–651, 2017.

[8] J. Almazan, A. Gordo, A. Fornes, and E. Valveny. Word spotting and recognition with embedded attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36 (12), pages 2552–2566, 2014.