Word Spotting in Scene Images based on Character Recognition

Dena Bazazian, Dimosthenis Karatzas
CVC, Universitat Autònoma de Barcelona
{dena.bazazian, dimos}@cvc.uab.es

Andrew D. Bagdanov
MICC, University of Florence
andrew.bagdanov@unifi.it

Abstract

In this paper we address the problem of unconstrained Word Spotting in scene images. We train a Fully Convolutional Network to produce heatmaps of all the character classes. Then, we employ the Text Proposals approach and, via a rectangle classifier, detect the most likely rectangle for each query word based on the character attribute maps. We evaluate the proposed method on ICDAR2015 and show that it is capable of identifying and recognizing query words in natural scene images.

1. Introduction

Reading text in the wild is important as it carries semantically rich information which can be employed in applications such as scene understanding and visual assistance. In this work we address the challenge of spotting text in scene images without restricting the words to a fixed lexicon or dictionary. Words which are typically out of dictionary include, for instance, cases where exclamation or other punctuation marks are present in words, telephone numbers, URLs, dates, etc. To this end, we train a Fully Convolutional Network (FCN) inspired by [13] to recognize individual characters (including letters, numbers, and punctuation) in scene images. Moreover, we detect text proposal regions in images using the approach proposed in [2]. Afterwards, we search for the query word in images by training a rectangle classifier to find the correspondence between the character attribute map given by the FCN on proposed rectangles and the character histogram of the query word. The key advantage of the proposed method is that it allows unconstrained out-of-dictionary word spotting independent from any dictionary or lexicon.

The contributions of this paper are the following. First, we propose a novel mid-level representation of the image in terms of character attribute maps by means of a Fully Convolutional Network. Second, we propose a novel pipeline that fuses the FCN produced representations with text proposals and the PHOC representation [1] to efficiently perform word spotting.

The remainder of the article is organized as follows. In the next section we review work related to our approach. In Section 3 we describe our proposed approach. We report our experimental results in Section 4, and draw conclusions in Section 5.

2. Related Work

In the past few years, scene text detection and recognition have been widely studied and significant progress has been achieved. Deep Convolutional Neural Networks (DCNNs) have become the standard approach for many computer vision tasks and DCNN methods are also state-of-the-art for text recognition. The authors of [8] looked at the problem of unconstrained text recognition by using generic object proposals and a CNN to recognize words from an extensive lexicon. However, the generic object proposal approach does not perform well on text detection tasks. The Text Proposals approach [5] introduced a text-specific object proposal method that is based on generating a hierarchy of word hypotheses according to the similarity region grouping algorithm. Later, the authors of [3] fused the Text Proposals technique with a Fully Convolutional Network (FCN) [13] in order to achieve high text region recall while considering significantly fewer candidate regions, while in follow-up work [2] they improve the pipeline to increase the speed of the text proposal generator. This approach has been applied to compressed images [4]. TextBoxes [11] repurposed the SSD detector [12] for word-wise text localization. Exploiting the robustness of SSD detector [12], the authors of [7] proposed an attention mechanism that directly detects the word-level bounding box.

In this work, we employ a character recognizer network, in contrast to [10] which applied a text recognition network. The key advantage of our approach is to learn characters individually and independently of a lexicon of words.

3. Proposed Method

In this section we introduce our proposed framework which is illustrated in Fig. 1.

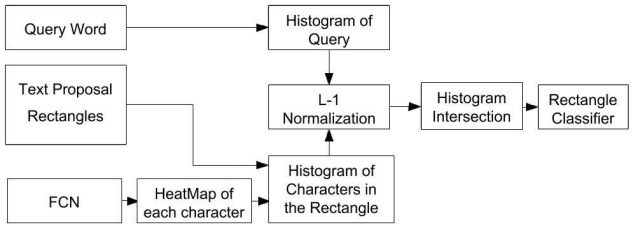

Figure 1. Diagram of Word spotting via Character detection.

3.1. Character recognizer

For the task of character recognition in scene images we use a Fully Convolutional Network (FCN) inspired by [13]. We train it on a synthetic dataset [6] to recognize and localize 38 character classes: background, 26 letters (case-insensitive), 10 digits, and one class for all punctuation symbols. Given an input image this FCN produces a heatmap with 38 channels, where each channel represents the probability of each character class at every location in the image.

3.2. Rectangle Classifier

For each image we also generate a list of ranked text proposal rectangles using the approach described in [2]. The FCN responses inside each text proposal region are pooled using integral images for each of the 38 channels of the heatmap. This yields a 38-dimensional vector for each of the text proposals, which we call the character *energy vector* of the proposal rectangle.

Given a list of query words, for each query we compute its histogram over the 38 character classes used to train our FCN. If there are n rectangles, the energy vector matrix (e) has a size of $n \times 38$ and the query histogram (q) is 1×38, as shown in Fig. 2.

Query words for spotting are represented as L1-Normalized histograms:

$$\overline{q} = \frac{q_i}{\sum_{i=1}^{38} q_i} \tag{1}$$

For each text proposal rectangle the normalized energy vector is given by:

$$\overline{e} = \frac{e_i}{\sum_{i=1}^{38} e_i} \tag{2}$$

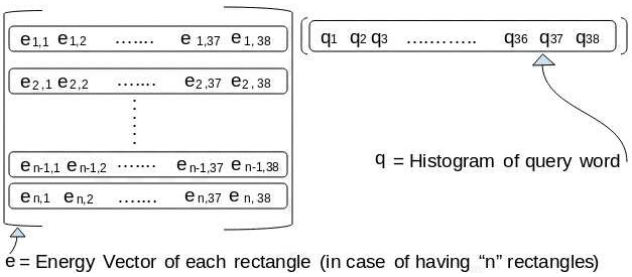

Figure 2. Energy vector and query histogram matrices.

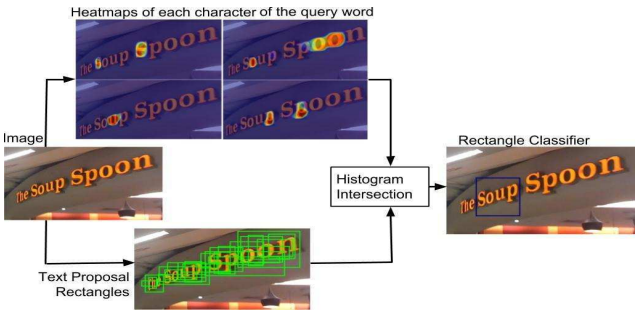

Figure 3. Word spotting via character detection.

To rank each proposal for each query word, we perform the Histogram Intersection of the normalized energy vector ($\overline{e}$) for n rectangle proposals and the normalized query histogram ($\overline{q}$) of the query word:

$$K(\overline{e}, \overline{q}) = \frac{\sum_{j=1}^{n} \min(\overline{e}_j, \overline{q})}{\min(\sum_{j=1}^{n} \overline{e}_j, \overline{q})} \tag{3}$$

Finally, we define the best matched text proposal rectangle with the query word by considering the highest K as shown in Fig. 3.

3.3. Improving the pipeline

We augmented our proposed method with two techniques: one to mitigate false positives due to background clutter, and another to improve the discriminative power of our representation.

3.3.1 Fusing text detector with character recognizer

In scene images, especially with cluttered backgrounds, many features in the background are often confused with characters. To address this problem, we trained the same FCN for two classes of text and non-text. Afterwards, we fuse the energy vector of these two nets together (e^{text} and e^{char}). For the background channel we compute the $\widehat{e_1}$ as:

$$\widehat{e_1} = e_1^{text} * e_1^{char}, \tag{4}$$

and for the character channels we compute $\widehat{e_i}$ as:

$$\widehat{e_i} = e_2^{text} * e_i^{char} \quad (where \quad i = 2:38). \tag{5}$$

In Fig. 4 (left) we illustrate the improvement from fusing two nets of text detector and character recognizer in the proposed word classifier.

3.3.2 Pyramidal Histogram Of Characters (PHOC)

To consider the order of characters in each query word, we have applied the pyramidal version of Histogram Of Characters (PHOC) inspired from [1]. With the naive approach described before, words such as "listen" and "silent" share the same histogram of characters representation. The

Figure 4. Improved results. Blue and red rectangles are with and without improvements, respectively. Left: combining the character recognizer and text detector networks: the error in the red rectangle is due to the strong response of "A" (query word is "ALDO"). Right: improvement from PHOC (query word is "cook").

Figure 5. Results of our proposed word spotting technique: green rectangles are our estimates, red rectangles are ground truth.

PHOC approach [1] improves the accuracy of our word spotting technique by increasing its discriminative power. Instead of computing a single histogram over characters for each word, we focus on different regions of the word. In our experiments we applied a two-level PHOC, but higher-level representations are straightforward. Hence, we compute histograms of characters over the first and second halves of the query word. Accordingly, we perform the Histogram Intersection of each half of the query word with each half of the text proposal. In Fig 4 (right) we illustrate the improvement of applying PHOC to the word classifier.

4. Experimental results

We evaluated our proposed method on the ICDAR2015 [9] End-to-End Task. We considered the Strongly Contextualised list of query words which consists of 100 words per each image, including all the words that appear in the image and a number of distractor words. Preliminary results yield an F-score of 38.2%. Moreover, some qualitative results of our experiments are shown in Fig. 5.

5. Conclusions and future work

In this work we proposed a technique addressing the problem of unconstrained word spotting for scene images based on character recognition. We extract character attribute maps of each query word using heatmaps from a Fully Convolutional Network. Afterwards, using a rectangle classifier that fuses FCN heatmaps with Text Proposals, we detect the most likely rectangle for each query word. We evaluated our proposed method on ICDAR2015 and it is capable of extracting query words. In the future we plan to incorporate contextual language models in our framework in order to go beyond the character-level recognition.

Acknowledgments

This work was supported by the project TIN2017-89779-P.

References

[1] 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.

[2] 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*, 2016.

[3] D. Bazazian, R. Gomez, A. Nicolaou, L. Gomez, D. Karatzas, and A. Bagdanov. Improving text proposals for scene images with fully convolutional networks. In *DLPR, ICPR*, 2016.

[4] L. Galteri, D. Bazazian, L. Seidenari, A. B. M. Bertini, A. Nicolaou, D. Karatzas, and A. Bimbo. Reading text in the wild from compressed images. In *Proc. ICCV*, pages 2399–2407, 2017.

[5] L. Gomez and D. Karatzas. Textproposals: a text-specific selective search algorithm for word spotting in the wild. *Pattern Recognition*, (70):60–74, 2017.

[6] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *Proc. CVPR*, pages 2315–2324, 2016.

[7] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li. Single shot text detector with regional attention. In *Proc. ICCV*, pages 3066–3074, 2017.

[8] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. *IJCV*, 116(1):1–20, 2016.

[9] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny. Icdar 2015 robust reading competition. In *Proc. ICDAR*, pages 1156–1160, 2015.

[10] H. Li, P. Wang, and C. Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. In *Proc. ICCV*, pages 5238–5246, 2017.

[11] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. In *Proc. AAAI*, pages 4161–4167, 2017.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. Ssd: Single shot multibox detector. In *Proc. ECCV*, pages 21–37, 2016.

[13] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. on PAMI*, 2016.