COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images

Andreas Veit[1,2], Tomáš Matera[2], Lukáš Neumann[3], Jiří Matas[3], Serge Belongie[1,2]
[1] Department of Computer Science, Cornell University [2] Cornell Tech
[3] Department of Cybernetics, Czech Technical University, Prague
[1]{av443,sjb344}@cornell.edu, [2]tomas@matera.cz, [3]{neumalu1,matas}@cmp.felk.cvut.cz

Abstract

This paper describes the COCO-Text dataset. In recent years large-scale datasets like SUN and Imagenet drove the advancement of scene understanding and object recognition. The goal of COCO-Text is to advance state-of-the-art in text detection and recognition in natural images. The dataset is based on the MS COCO dataset, which contains images of complex everyday scenes. To reflect the diversity of text in natural scenes, we annotate text with (a) location in terms of a bounding box, (b) fine-grained classification into machine printed text and handwritten text, (c) classification into legible and illegible text, (d) script of the text and (e) transcriptions of legible text. The dataset contains over 173k text annotations in over 63k images.

1. Introduction

To advance the understanding of text in unconstrained scenes we present a new large-scale dataset for text in natural images.[1] The dataset is based on the Microsoft COCO dataset [2] that annotates common objects in their natural contexts. Combining rich text annotations and object annotations in natural images provides a great opportunity for research in scene understanding as well as text detection and recognition. Combining text with object annotations allows for contextual reasoning about scene text and objects. During a pilot study of state-of-the-art photo OCR methods on MS COCO, we made two key observations: First, text in natural scenes is very diverse ranging from legible machine printed text on street signs to illegible graffiti and handwritten notes. Second, while the domain of scene text detection and recognition enjoys significant advances in recent years, there is still way to go to reach the performance needed for real world applications. Figure 1 shows sample images from the dataset illustrating the diversity of scene text in natural images and the challenges for text detection and recognition. The dataset contains 63,686 images with 173,589 labeled text regions. This is in stark contrast to the scale of current datasets, with mostly iconic text and evaluation sets containing hundreds of images at best. For each text region, we provide the location in terms of bounding boxes, classifications in terms of legibility, category (*e.g.* machine printed or hand written) and script of the text, as well as transcriptions in case of legible text with western script. A detailed description of our annotation pipeline can be found on the project website.[1] In addition, we analyse three leading state-of-the-art photo OCR algorithms on our dataset. Some methods achieve good detection precision and transcription accuracy. However, recall for text detection is considerably degraded. In particular, for illegible text none of the methods shows viable functionality. These significant shortcomings motivate future work.

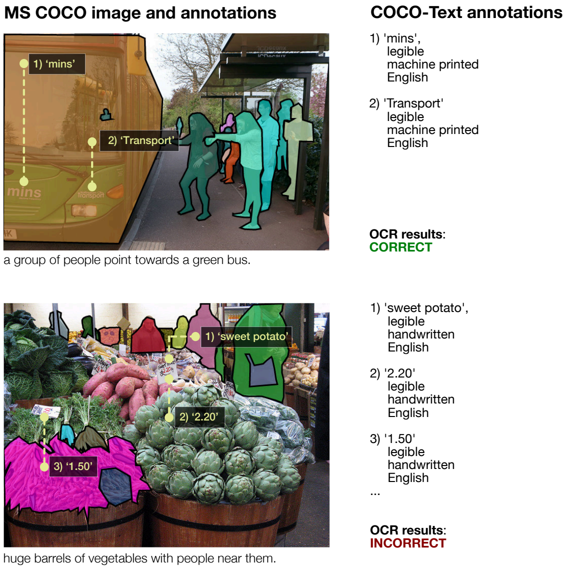
MS COCO image and annotations
COCO-Text annotations
1) 'mins'
2) 'Transport'
mins
1) 'mins', legible machine printed English
2) 'Transport' legible machine printed English
a group of people point towards a green bus.
OCR results: CORRECT
1) 'sweet potato'
2) '2.20'
3) '1.50'
1) 'sweet potato', legible handwritten English
2) '2.20' legible handwritten English
3) '1.50' legible handwritten English
...
huge barrels of vegetables with people near them.
OCR results: INCORRECT

Figure 1. Left: Example MS COCO images with object segmentation and captions. Right: COCO-Text annotations. For the top image, the photo OCR finds and recognizes the text printed on the bus. For the bottom image, the OCR does not recognize the handwritten price tags on the fruit stand.

[1] available at http://vision.cornell.edu/se3/coco-text

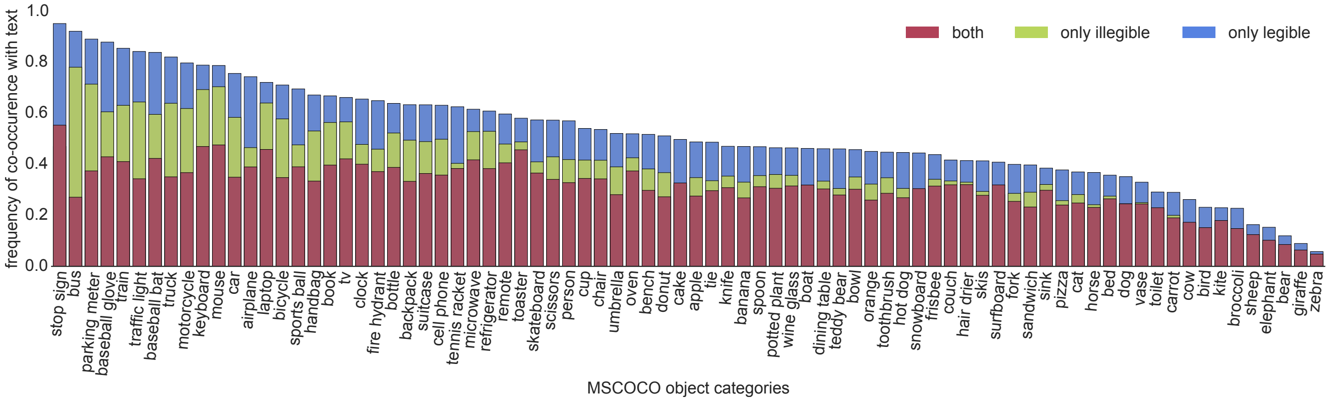

Figure 2. Frequency that objects in MS COCO co-occur with text. It can be seen that the presence of certain objects is very informative regarding text presence. Especially traffic and sports scenes almost always contain text and nature scenes with animals rarely contain text.

2. Dataset Statistics

We analyze COCO-Text and compare to other popular scene text datasets, in particular, ICDAR 03 and 15 [1]. The first difference is that images of COCO-Text were not collected with text in mind. This leads to generally more natural text. As a consequence the spatial distribution of text in the images is wider than in related datasets. Further, COCO-Text is the only scene text datset containing images without any text. In particular, 50% of the images do not contain text. Overall, there are in average 2.73 instances of text per image. Considering only images with text, the average is 5.46. Further, text instances are annotated with rich attributes. In addition to location and transcriptions, COCO-Text also annotates text legibility, type of the text as well as the script of the text. Overall, 60.3% of text is legible and 39.7% illegible. The majority of text is machine printed with 86.4%. Only 4.6% of text is handwritten and 9% is borderline or from other not captured categories. Another key aspect of COCO-Text is its contextual information. It is part of the larger context of MSCOCO and thus enables contextual reasoning between scene text and objects. This is relevant as context is highly informative for many applications. Figure 2 shows the frequency of co-occurrences of MSCOCO object categories with scene text. It shows that the presence of certain objects is very informative regarding text presence. Lastly, COCO-Text has a larger scale than related datasets. With 63,686 images and 173,589 text annotation, it is more than 14 times larger than ICDAR 15.

3. Algorithmic Analysis

We follow the evaluation scheme as used in the ICDAR robust reading competition for end-to-end recognition of *incidental scene text*. Detailed instructions can be found in the extended version on the project website. We take three state-of-the-art photo OCR algorithms from our collaborators at Google, TextSpotter and VGG and evaluate their detection, transcription and end-to-end text spotting results on our dataset. Results are anonymized. On the positive side, methods A and B have good detection precision with 79.82 and 76.38% respectively. Further, we observe good recognition accuracy, with the best method A achieving 83.76%. However, detection performance is very weak overall. While method A finds considerable amounts of legible machine printed text with 34.99%, no method performs satisfactory. Even lower results are observed on legible handwritten text. Lastly, no method has even viable functionality to find illegible text.

4. Discussion

We introduced COCO-Text, the first large-scale dataset for detecting and recognizing text in natural images and also the first dataset to annotate scene text with attributes such as legibility and type of text. We further evaluate state-of-the-art photo OCR algorithms on our dataset. While the results indicate satisfactory precision, we identify significant shortcomings especially for detection recall.

Acknowledgments

Funding for the crowd worker tasks was provided by a Microsoft Research Award. We like to thank our collaborators Kai Wang from Google and Ankush Gupta and Andrew Zisserman from Oxford for providing photo OCR output for our data collection effort and for valuable feedback throughout defining and collecting the dataset. A.V. was supported by the AOL-Program for Connected Experiences.

References

[1] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al. ICDAR 2015 competition on Robust Reading. In *ICDAR '15*. IEEE.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV '14*. Springer.