Text Detection in Stores from a Wearable Camera Using a Repetition Prior

Bo Xiong Kristen Grauman
University of Texas at Austin
{bxiong,grauman} @cs.utexas.edu

1. Introduction

Text *detection* in natural scenes requires localizing regions in the image containing text—no matter what that text says, or what font it is written in. Text, signs, and labels are ubiquitous and informative in many natural environments. As such, with the increasing use of portable mobile and wearable computing platforms, reliable text detection is critical for many applications. For example, text detection (and a subsequent recognition process) is vital to real-world applications such as sign reading for place localization for tourists or mobile robots; for assistive technology to help visually impaired users navigate the world with more independence; and for image/video indexing and retrieval based on scene text or graphical overlays.

Whereas early work focused on constrained scenarios, such as finding lines of text in a document, today's methods tackle text detection "in the wild" in natural scenes. Doing so requires robustness to different fonts, languages, illuminations, orientations, occlusions, and clutter.

Despite a surge of exciting progress in natural scene text detection, we observe that a domain of great practical interest—stores—has largely been ignored. Current methods and datasets often focus on outdoor StreetView-style settings where text may appear on storefront signs, street signs, addresses, or license plates. However, text is also abundant in indoor store environments, where text appears on the labels of products that line the shelves (e.g., grocery stores, bookstores, electronics, etc.). Detecting it would assist in identifying products, retrieving relevant product reviews, reading prices or helping a visually impaired user complete his shopping list. Such applications with wearable computing platforms promise to revolutionize the traditional shopping experience, mixing the bricks-and- mortar environment with the online marketplace—particularly for wearable cameras.

However, text detection in a store presents its own challenges. Images of store shelves contain many products crowded together. Even worse, many products contain design patterns that share similar texture as text, and most have a high density of text occurrences. These properties can be problematic for mainstream text detection systems using sliding window or connected components. Furthermore, whereas existing datasets often contain images purposely taken so the text is somewhat prominent within the view, imagery captured more casually and even passively (i.e., on a shopper's wearable camera) will lack helpful cues implicit in the image composition. See Figure 1.

Figure 1. There are distinct challenges in finding text in store settings (right) versus street-side scenes (left). We propose to exploit the fact that products appear in duplicate on store shelves when performing text detection.

We propose an approach to text detection that specifically targets indoor store settings. As discussed above, the high density of products on a shelf creates many nuisances for detecting text. However, that same density comes along with one helpful factor: *each product typically appears multiple times on the shelf, side by side*. Our key insight is that duplicate occurrences of text can be a valuable prior for a text detector. Intuitively, a detector primed to see multiple instances of the same word can prioritize windows that have repeated support. We call this a *repetition prior*. Please see our full WACV 2016 paper [6] and project page [1] for more algorithmic details and results.

2. Approach

Our goal is text detection in store environments. The input is an image, the output is a set of confidence-ranked bounding boxes believed to contain one word of text each.

Our "repetition prior" has value in settings where products repeat on the shelves, and they have labels with texts and/or visible price tags. This applies to places like grocery stores, bookstores, music stores, etc., but not arbitrary

[1] http://vision.cs.utexas.edu/projects/text_repetition

Figure 2. Example text detections for our method and the best baseline [4]. We stress that we show only the top-ranked detections for clearest visualization; additional boxes with lower confidence are also found in each image. Best viewed on pdf with zoom.

stores. A user may snap a photo of the shelves using a mobile device or simply walk down the aisle wearing a camera like Google Glass; we study both such scenarios in our experiments. We assume no prior knowledge about a lexicon nor any prior knowledge about how many unique products appear in a single input image.

Given an image, we first generate candidate text boxes using the the state-of-the art method of Jaderberg et al. [4], and then identify similar candidate boxes. We pose the problem of finding multiple occurrences of the same text as finding connected components in a graph. We build a graph for the image, where each node is a proposal box. To define adjacency between the nodes, we consider three criteria: visual similarity, size, and overlap. Next, for each identified cluster, we expand the recall rate by matching a representative of each cluster to other visually similar regions that were ignored by the initial text detector. Finally, based on the results of the clustering and matching stages, we rank each text bounding box with a confidence score.

3. Results

We evaluate our approach on two challenging datasets and compare to multiple recent text detection methods. We consider two realistic datasets containing grocery store images: GROCERY PRODUCTS [2] and GLASS VIDEO [5] taken with Google Glass.

Comparison to existing methods We compare to three recent methods for lexicon-free text detection: 1) STROKE WIDTH TRANSFORM [1]; 2) MSER TEXT DETECTION [3]; 3) DEEP TEXT SPOTTING [4].

Figure 3 shows the precision-recall results for each dataset. Overall, our method outperforms the existing methods. Our gains over the two non-learning approaches ([1, 3]) are largest, reinforcing recent findings about the power of learned character detectors that leverage large training data sets. Furthermore, we see sizeable gains over the state-of-the-art deep learning approach [4], particularly in terms of precision. This is an important empirical finding, since our method specifically builds on the output of [4], enhancing it with the repetition prior.

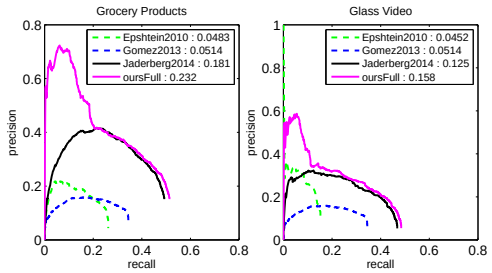

Figure 3. Text detection accuracy on the GROCERY PRODUCTS (left) and GLASS VIDEO (right) dataset. Our method improves the results of a state-of-the-art text detector [4].

Qualitative examples Next, we present example text detections in Figure 2.The first rows is our method and the second row is the best competing baseline [4].

These image examples help illustrate where and how our repetition prior helps. For example, in the leftmost image of the first row, our method is able to find three repeating words and considers them more confident than the non-repeating candidates. These identified repeating text candidates are in fact true text and are properly localized. On the other hand, the baseline misclassifies part of a product shelf as text, possibly due to its similar appearance with the letter i or l. Such non-text regions are less likely to repeat, and therefore our prior helps disregard that error. Overall, our method focuses attention on valid repeating texts and can ignore spurious proposals.

References

[1] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *CVPR*, 2010.

[2] M. George and C. Floerkemeier. Recognizing products: A per-exemplar multi-label image classification approach. In *ECCV*. 2014.

[3] L. Gómez and D. Karatzas. Multi-script text extraction from natural scenes. In *ICDAR*, 2013.

[4] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *ECCV*. 2014.

[5] S. Rallapalli, A. Ganesan, K. Chintalapudi, V. N. Padmanabhan, and L. Qiu. Enabling physical analytics in retail stores using smart glasses. In *ACM MobiCom*, 2014.

[6] B. Xiong and K. Grauman. Text detection in stores using a repetition prior. In *WACV*, 2016.