Computer Vision and Image Understanding 145 (2016) 30–42

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Enhancing energy minimization framework for scene text recognition with top-down cues☆

Anand Mishra[a,*], Karteek Alahari[b], C.V. Jawahar[a]

[a] Center for Visual Information Technology, IIIT Hyderabad, India
[b] THOTH Team, Inria Grenoble Rhone-Alpes, Laboratoire Jean Kuntzmann, CNRS, Université Grenoble Alpes, France

ARTICLE INFO

Article history:
Received 4 April 2015
Accepted 4 January 2016
Available online 21 January 2016

Keywords:
Scene text understanding
Text recognition
Lexicon priors
Character recognition
Random field models

ABSTRACT

Recognizing scene text is a challenging problem, even more so than the recognition of scanned documents. This problem has gained significant attention from the computer vision community in recent years, and several methods based on energy minimization frameworks and deep learning approaches have been proposed. In this work, we focus on the energy minimization framework and propose a model that exploits both bottom-up and top-down cues for recognizing cropped words extracted from street images. The bottom-up cues are derived from individual character detections from an image. We build a conditional random field model on these detections to jointly model the strength of the detections and the interactions between them. These interactions are top-down cues obtained from a lexicon-based prior, i.e., language statistics. The optimal word represented by the text image is obtained by minimizing the energy function corresponding to the random field model. We evaluate our proposed algorithm extensively on a number of cropped scene text benchmark datasets, namely Street View Text, ICDAR 2003, 2011 and 2013 datasets, and IIIT 5K-word, and show better performance than comparable methods. We perform a rigorous analysis of all the steps in our approach and analyze the results. We also show that state-of-the-art convolutional neural network features can be integrated in our framework to further improve the recognition performance.

1. Introduction

The problem of understanding scenes semantically has been one of the challenging goals in computer vision for many decades. It has gained considerable attention over the past few years, in particular, in the context of street scenes [1–3]. This problem has manifested itself in various forms, namely, object detection [4,5], object recognition and segmentation [6,7]. There have also been significant attempts at addressing all these tasks jointly [2,8,9]. Although these approaches interpret most of the scene successfully, regions containing text are overlooked. As an example, consider an image of a typical street scene taken from Google Street View in Fig. 1. One of the first things we notice in this scene is the sign board and the text it contains. However, popular recognition methods ignore the text, and identify other objects such as car, person, tree, and regions such as road, sky. The importance of text in images is also highlighted in the experimental study conducted by Judd et al. [10]. They found that viewers fixate on text when shown images containing text and other objects. This is further evidence that text recognition forms a useful component in understanding scenes.

In addition to being an important component of scene understanding, scene text recognition has many potential applications, such as image retrieval, auto navigation, scene text to speech systems, developing apps for visually impaired people [13,14]. Our method for solving this task is inspired by the many advancements made in the object detection and recognition problems [4,5,7,15]. We present a framework for recognizing text that exploits bottom-up and top-down cues. The bottom-up cues are derived from individual character detections from an image. Naturally, these windows contain true as well as false positive detections of characters. We build a conditional random field (CRF) model [16] on these detections to determine not only the true positive detections, but also the word they represent jointly. We impose top-down cues obtained from a lexicon-based prior, i.e., language statistics, on the model. In addition to disambiguating between characters, this prior also helps us in recognizing words.

The first contribution of this work is a joint framework with seamless integration of multiple cues – individual character

☆ This paper has been recommended for acceptance by Daniel Lopresti.

* Corresponding author. Fax: +91 40 6653 1413.
E-mail address: anand.mishra@research.iiit.ac.in, anandmishra.jsp@gmail.com (A. Mishra).

http://dx.doi.org/10.1016/j.cviu.2016.01.002
1077-3142/

Fig. 1. A typical street scene image taken from Google Street View. It contains very prominent sign boards with text on the building and its windows. It also contains objects such as car, person, tree, and regions such as road, sky. Many scene understanding methods recognize these objects and regions in the image successfully, but overlook the text on the sign board, which contains rich, useful information. The goal of this work is to address this gap in understanding scenes.

detections and their spatial arrangements, pairwise lexicon priors, and higher-order priors – into a CRF framework which can be optimized effectively. The proposed method performs significantly better than other related energy minimization based methods for scene text recognition. Our second contribution is devising a cropped word recognition framework which is applicable not only to closed vocabulary text recognition (where a small lexicon containing the ground truth word is provided with each image), but also to a more general setting of the problem, i.e., open vocabulary scene text recognition (where the ground truth word may or may not belong to a generic large lexicon or the English dictionary). The third contribution is comprehensive experimental evaluation, in contrast to many recent works, which either consider a subset of benchmark datasets or are limited to the closed vocabulary setting. We evaluate on a number of cropped word datasets (ICDAR 2003, 2011 and 2013 [17], SVT [18], and IIIT 5K-word [19]) and show results in closed and open vocabulary settings. Additionally, we analyzed the effectiveness of individual components of the framework, the influence of parameter settings, and the use of convolutional neural network (CNN) based features [20].

The rest of the paper is organized as follows. In Section 2 we discuss related work. Section 3 describes our scene text recognition model and its components. We then present the evaluation protocols and the datasets used in experimental analysis in Section 4. Comparison with related approaches is shown in Section 5, along with implementation details. We then make concluding remarks in Section 6.

Table 1
Our IIIT 5K-word dataset contains a few less challenging (easy) and many very challenging (hard) images. To present analysis of the dataset, we manually divided the words in the training and test sets into *easy* and *hard* categories based on their visual appearance. The recognition accuracy of a state-of-the-art commercial OCR – ABBYY9.0 – for this dataset is shown in the last column. Here, we also show the total number of characters, whose annotations are also provided, in the dataset.

	#words	#characters	ABBYY9.0 (%)
	Training set		
Easy	658	–	44.98
Hard	1342	–	16.57
Total	2000	9658	20.25
	Test set		
Easy	734	–	44.96
Hard	2266	–	5.00
Total	3000	15269	14.60

2. Related work

The task of understanding scene text has gained a huge interest for more than a decade [11,12,20–31]. It is closely related to the problem of Optical Character Recognition (OCR), which has a long history in the computer vision and pattern recognition communities [32]. However, the success of OCR systems is largely restricted to text from scanned documents. Scene text exhibits a large variability in appearance, as shown in Fig. 2, and can prove to be challenging even for the state-of-the-art OCR methods (see Table 1 and [11,12]). The problems in this context are: (1) text localization, (2) cropped word recognition, and (3) isolated character recognition. They have been tackled either individually [21,27,33], or jointly [11,20,23,29]. This paper focuses on addressing the cropped word recognition problem. In other words, given an image region (e.g., in the form of a bounding box) containing text, the task is to recognize this content. The core components of a typical cropped word recognition framework are: localize the characters, recognize them, and use statistical language models to compose the characters into words. Our framework builds on these components, but differs from previous work in several ways. In the following, we review the prior art and highlight these differences. The reader is encouraged to refer [34] for a more comprehensive survey of scene text recognition methods.

A popular technique for localizing characters in an OCR system is to binarize the image and determine the potential character locations based on connected components [35]. Such techniques have also been adapted for scene text recognition [12], although with limited success. This is mainly because obtaining a clean binary output for scene text images is often challenging; see Fig. 3 for examples. An alternative approach is proposed in [36] using gradient information to find potential character locations. More recently, Yao et al. [31] proposed a mid-level feature based technique

Fig. 2. Challenges in scene text recognition. A few sample images from the SVT and IIIT 5K-word datasets are shown to highlight the variation in view point, orientation, non-uniform background, non-standard font styles and also issues such as occlusion, noise, and inconsistent lighting. Standard OCRs perform poorly on these datasets (as seen in Table 1 and [11,12]).