Pattern Recognition 58 (2016) 204–215

Contents lists available at ScienceDirect

Pattern Recognition

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

Could scene context be beneficial for scene text detection?

Anna Zhu [a,*], Renwu Gao [b], Seiichi Uchida [b]

[a] Huazhong University of Science and Technology, State Key Lab for Multispectral Information Processing Technology, School of automation, Wuhan, China
[b] Kyushu University, Human Interface Laboratory, Information Science and Electrical Engineering, Fukuoka, Japan

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form
24 March 2016
Accepted 15 April 2016
Available online 26 April 2016

Keywords:
Scene text detection
Fully connected CRF
Convolutional neural network
Character feature
Context feature

ABSTRACT

Scene text detection and scene segmentation are meaningful tasks in the computer vision field. Could the semantic scene segmentation assist scene text detection in any degree? For example, can we expect the probability of a region being text is low if its surrounding segment, i.e., its context, is labeled as sky? In this paper, we have a positive answer by constructing a scene context-based text detection model. In this model, we use texton features and a fully-connected conditional random field (CRF) to estimate pixel-level scene class's probability to be considered as image's context feature. Meanwhile, maximally stable extremal regions (MSERs) are extracted, integrated and extended as image patches of character candidates. Then, each image patch is fed to a simple two-layer convolutional neural network (CNN) to automatically extract its character feature. The averaged context feature of the corresponding patch is considered as the patch's context feature. The character feature and context feature are fused as the input into a support vector machine for text/non-text determination. Finally, as a post-processing, neighboring text regions are grouped hierarchically. The performance evaluation on ICDAR2013 and SVT databases, as well as a preliminary evaluation on a patch-level database, proves that the scene context can improve the performance of scene text detection. Moreover, the comparative study with state-of-the-art methods shows the top-level performance of our method.

1. Introduction

Recently, scene text reading captures much attention in the computer vision field [1,2]. It refers to an attempt to recognize text from camera-captured images and contains two parts: scene text detection and scene text recognition. The technologies of this area can be applied to various applications, such as language translation system, visual-based navigation, and content-based web image search. Accurate detection of text in natural scene images is an essential step and primary work for successfully recognizing scene text. Therefore, we focus our attention on scene text detection in this paper.

Scene text detection meets great challenge [3,4]. The difficulty comes from the huge variations of scene text. Specifically, scene text (1) varies in size, color, font, and style; (2) has no clear layout; (3) tends to have complex background, and (4) undergoes non-uniform lighting, partial occlusion, blur, rotation, perspective distortion, and low resolution.

To tackle with those difficulties, we propose a framework to detect scene text by means of scene context information. The idea arises from the observation that text appear frequently in certain scene context and rarely in others. For example, they are always embedded in signboard, cars, books and other surfaces, but rarely appear in rivers, the sky, trees or grass. If this idea is valid, a better detection performance can be expected by using the framework shown in Fig. 1.

In this framework, we use not only the character feature but also context feature for discriminating a connected component (CC) into text or non-text. In this situation, even if the character feature suggests that a non-text CC is a character or a part of a character, it may still be discriminated into non-text, when the context feature suggests its surrounding segment is “sky”. Conversely, even if the character feature suggests that a text CC is non-text, it can be detected as text, when the context feature suggests its surrounding segment is “signboard”.

Scene context information represents the scene class attribution of pixels in the images, which can be extracted through semantic scene segmentation. Though semantic scene segmentation has been another difficult task, recent researches [5,6] show very promising results. In our approach, we employ Philipp's work [7], in which images are segmented and semantic labeled in pixel-level by TextonBoost and then refined by a fully-connected conditional random field (CRF). From the segmentation result, we obtain K-dimensional feature vector of each pixel, namely the scene context, where K is the number of scene classes (such as sky and signboard) and the kth element of the vector is a probability

* Corresponding author. Tel.: +86 186 2770 8137.
E-mail address: annakkk@live.com (A. Zhu).

http://dx.doi.org/10.1016/j.patcog.2016.04.011
0031-3203/

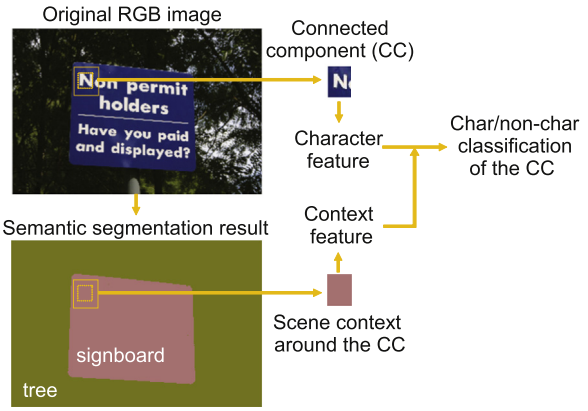

Fig. 1. Our hypothesis: confuse context information for classification.

that the pixel lies on the *k*th scene class. The averaged scene context probabilities of a CC region is considered as this CC's context feature. The context feature is then combined with the character feature generated by a two-layer convolutional neural network (CNN) [8] and fed to a support vector machine (SVM) classifier for text/non-text classification.

The contribution of our work is threefold. The first contribution is to propose a new idea of utilizing scene context for scene text detection. In traditional scene text detection researches, they focus only on "text-ness" at each region in a scene image. Specifically, they assume several heuristics about texts, such as, high contrast, high spatial frequency, dense edges, uniform stroke width, and then use them for detection. Even though these heuristics work reasonably, they cannot avoid false alarms and false negatives. False alarms will be detected around text-like shapes in scene. For example, car tires can be detected as "o" because it satisfies most of the above heuristics. False negatives will happen on texts with irregular appearance. For example, texts printed in a fancy font may not have uniform stroke width and texts on complex background may not have enough contrast. Some recent methods are free from those heuristics by using machine learning techniques, but they learn text-ness only and thus still have similar false alarms and false negatives. The introduction of scene text will be an essential solution for this problem.

The second contribution is to emphasize the usefulness of semantic segmentation for scene text detection. Semantic segmentation is a rather new technology and developed mainly for scene object recognition and scene understanding. To our best knowledge, semantic segmentation has never been used for scene text detection, i.e., there is no class "text" in the task of semantic scene segmentation. This may come from that most semantic segmentation methods assume some smoothness and prevent detecting very fine and thin structures, i.e., texts. We will see that semantic segmentation is still very useful for scene text detection by the help of CC-based text detection strategy, even though semantic segmentation has no enough resolution to detect individual text by itself.

The third contribution of our work is that, to prove our hypothesis, four variable controlled comparison experiments, which control the variable whether to use context feature and how to train CNN, are performed to provide adequate and valid proof for test. Besides, a patch-level database is given in this work. In this database, all the patch images are manually labeled and cropped from text regions in non-overlapping images of the ICDAR 2011 and ICDAR 2013's training databases. The results on this patch-level database and ICDAR 2013 database achieve convincible performance.

The following section gives an overview of our pipeline. We review related work in two directions in Section 2. Section 3 presents our proposed method in detail. In Section 4, we give the experimental results which include the details of databases and the experimental setup. Finally, Section 5 gives a summarization and conclusion of this paper.

2. Related work

In this section, we summarize some related previous work. Since our work refers to scene segmentation but focus on scene text detection, we introduce the related works on scene segmentation briefly and text detection methods emphatically.

2.1. Scene segmentation methods

Scene image segmentation aims to label every pixel in the image with several predetermined classes, thus concurrently perform recognition and segmentation of multiple classes. The segmentation techniques can be classified as follows: graph-based approaches, region-based approaches, boundary detection approaches, perceptual organization approaches, multi-class image segmentation, and hybrid approaches.

The graph-based image segmentation approach defines the boundaries between regions by measuring the dissimilarity between the neighboring pixels by a graph, where the node representing each pixel and the weights denoting the dissimilarity between pixels [9]. Region-based techniques make use of common patterns in intensity values within a cluster of neighboring pixels and group regions according to their anatomical or functional roles [10]. The boundary detection approach refers to a contour in the image plane that represents dissimilar pixels between the neighboring segments [11]. Perceptual organization refers to a basic capability of the human visual system to obtain relevant groupings and structures from an image without having prior knowledge of the image's contents [12]. Multi-class image segmentation uses one of a number of classes (e.g., road, sky and water, etc) for labeling every pixel in an image. Many state-of-the-art methods first over-segment the image into superpixels (or small coherent regions) and then classify each region [13,14]. Hybrid techniques combine the above segmentation approaches [15] to take advantages of them. Our scene segmentation method belongs to hybrid method, because it combines multi-class image segmentation methods and graph-based methods together.

2.2. Scene text detection methods

Scene text detection involves extracting regions that contain text and filtering out regions without text in the images. It is a typical classification problem. We classify the text detection methods into three categories: global-classification methods, local-classification methods, and hybrid methods.

The global-classification methods split images to region-based patches and attempt to classify these patches into text/non-text by using feature, such as Histogram of Oriented Gradients (HOG) [16], Local Binary Pattern (LBP) [17], and learns classifiers [18,19] on these features to discriminate text and non-text. Those patches are selected by scanning the whole image with sliding windows. To resist different sizes of text, multi-scale strategy are typically used. After classification, the patches containing text are merged into text blocks. Chen and Yuille [20] trained a cascade of 4 strong AdaBoost classifiers containing 79 informative features to classify the regions. Wang et al. [21] classified all the patches extracted from the sliding window by using a random ferns classifier trained on HOG features. Neumann and Matas [22] used sliding windows for stroke detection by convolving the gradient field with a set of oriented bar filters. Mishra et al. [23] exploited bottom-up cues