Pattern Recognition 47 (2014) 2853–2866

Contents lists available at ScienceDirect

Pattern Recognition

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

End-to-end scene text recognition using tree-structured models

Cunzhao Shi, Chunheng Wang*, Baihua Xiao, Song Gao, Jinlong Hu

The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 20 June 2013
Received in revised form
27 January 2014
Accepted 23 March 2014
Available online 2 April 2014

Keywords:
End-to-end
Scene text recognition
Part-based tree-structured models (TSMs)
Normalized pictorial structure

ABSTRACT

Detecting and recognizing text in natural images are quite challenging and have received much attention from the computer vision community in recent years. In this paper, we propose a robust end-to-end scene text recognition method, which utilizes tree-structured character models and normalized pictorial structured word models. For each category of characters, we build a part-based tree-structured model (TSM) so as to make use of the character-specific structure information as well as the local appearance information. The TSM could detect each part of the character and recognize the unique structure as well, seamlessly combining character detection and recognition together. As the TSMs could accurately detect characters from complex background, for text localization, we apply TSMs for all the characters on the coarse text detection regions to eliminate the false positives and search the possible missing characters as well. While for word recognition, we propose a normalized pictorial structure (PS) framework to deal with the bias caused by words of different lengths. Experimental results on a range of challenging public datasets (ICDAR 2003, ICDAR 2011, SVT) demonstrate that the proposed method outperforms state-of-the-art methods both for text localization and word recognition.

1. Introduction

With the rapid growth of camera-based applications readily available on smart phones and portable devices, understanding the pictures taken by these devices semantically has gained increasing attention from the computer vision community in recent years. Among all the information contained in the image, text, which carries semantic information, could provide valuable cues about the content of the image and thus is very important for human as well as computer to understand the scenes.

Generally speaking, to understand the information carried by text in the image, we need to recognize the text detected from the images. Since text detection is the premise for recognition, a lot of methods have been proposed to address the problem of detecting and localizing text in scene images [1–9] and some have reported promising localization performance. For cropped character and word recognition, some previous methods [2,6], used off-the-shelf OCR for subsequent recognition, whose performance was unsatisfactory. Although many commercial OCR systems work well on scanned documents under controlled environment, they performed poorly on scene text images due to the unsatisfactory binarization results of text images of low quality. Some recent methods [10–14] propose to recognize scene text without binarization. After getting the raw detection/recognition results on the original image, conditional random fields [12,13] (CRF) or pictorial structures [10,11] are used to get the final recognition results.

End-to-end scene text recognition, namely detecting and also recognizing text from natural images, is our final goal. Although some end-to-end text recognition methods have been proposed recently, the performance is far from enough to put them into practice. The reasons are two-fold: (1) most of the previous text detection and recognition methods do not make use of the intrinsic global structure information of characters, which is essential to achieve good system performance especially for scene characters due to the unconstrained lighting conditions, fonts, deformations, occlusions, sometimes low resolution and complex background; and (2) most existing methods separate text detection and recognition stages which should be combined together to help each other.

In fact, characters are designed by humans and each character has unique structure representing itself. Therefore, no matter how the background changes or the character deteriorates, as long as the structure remains unchangeable, we could recognize them by detecting the unique structure from cluttered background. In other words, humans naturally make use of character-specific structure information when recognizing characters from scene images. Moreover, when humans try to recognize scene characters with distortions and complex background, the detection of the character from complex background and the recognition of the character are somehow interdependent. On one hand, the unique structure of each character helps us to detect the characters from

* Corresponding author. Tel./fax: +86 10 62650820.
E-mail address: chunheng.wang@ia.ac.cn (C. Wang).

http://dx.doi.org/10.1016/j.patcog.2014.03.023
0031-3203/

complex background and on the other hand, detecting the character-specific structure from complex background also helps us to recognize the character. In other words, humans naturally combine detection and recognition together when recognizing characters from scene images.

In this paper, we propose a novel end-to-end scene text recognition method using tree-structured character and word models. We use part-based tree-structure to model each category of characters so as to make use of the intrinsic global structure information as well as the local appearance information. The TSM could detect each part of the character and recognize the unique structure as well, seamlessly combining character detection and recognition together. As the TSMs could accurately detect characters from complex background, for text localization, we apply TSMs for all the characters on the candidate text regions to eliminate the false positives, and search the possible missing characters as well. While for word recognition, we propose a normalized pictorial structure framework to deal with the bias towards shorter words by adding the normalization term. Experimental results on a range of challenging public datasets (ICDAR 2003, ICDAR 2011, SVT) demonstrate that the proposed method outperforms state-of-the-art methods significantly both for text localization and word recognition.

The rest of the paper is organized as follows. Related word is reviewed in Section 2. An overview of the proposed method is given in Section 3. The text localization method is described in Section 4 and the normalized pictorial structure framework for scene text recognition is detailed in Section 5. Experimental results and discussions are given in Section 6 and conclusions are drawn in Section 7.

2. Related work

Most of the previous work mainly addresses subtasks of the end-to-end scene text recognition system–text detection [1–9], cropped character recognition [10,15–18] and word recognition [10–14]. Next, we will briefly review some related work on the subtasks and the end-to-end systems as well.

2.1. Text localization

Most of the existing methods of text detection could be roughly classified into two categories: region-based and connected component (CC) based. Region-based methods need to scan the image at multiple scales and use a text/non-text classifier to find the potential text regions. Chen and Yuille [2] proposed a fast text detector based on a cascade AdaBoost classifier. Coates et al. [18] proposed to learn features automatically from unlabeled data using unsupervised feature learning and then train a liner SVM to classify whether a sliding window is text or not.

As opposed to the region-based method, CC-based methods first use various approaches such as edge detection, color clustering or stroke width transform (SWT) to get the CCs, and heuristic rules or classifiers are used to remove non-text CCs. Pan et al. [7] adopted the region-based classifier to get the initial CCs and use CRF to filter non-text components. Epshtein et al. [5] used the CCs in a stroke width transformed image to form text line and the results on the ICDAR 2005 competition dataset [19] show promising performance. Shivakumara et al. [9] proposed to extract CCs by performing K-means clustering in the Fourier-Laplacian domain, and use text straightness and edge density to remove false positives. Chen et al. [20] employ edge-enhanced MSER as basic text candidates and geometric filtering as well as stroke width information is used to exclude non-text objects.

For region-based methods, considering the high degree of intraclass variation of scene characters and unpredictability of the possible background, it is quite difficult to train a perfect text/non-text classifier only using the statistical features. While for CC based methods, it is also difficult to design a reliable CC analyzer just using heuristic rules or statistical features to eliminate false positives without losing the text components.

2.2. Cropped word recognition

Most of the previous work on scene text recognition could be roughly classified into two categories: traditional Optical Character Recognition (OCR) based and object recognition based methods. For traditional OCR based methods, they focus on the binarization process which segments text from background and various approaches [21–27] have been proposed to binarize images with low quality or complex background. These methods could indeed improve the word recognition rates for a certain application. However, since text in natural images has unconstrained resolution, illumination condition, size and font style, quite part of the binarization results are still unsatisfactory. Moreover, the loss of information during the binarization process is almost irreversible, which means if the binarization result is poor, the chance of correctly recognizing the text is quite small.

On the other hand, object recognition based methods assume that scene character is quite similar to the generic object with a high degree of intraclass variation. These methods try to recognize scene text as a whole without binarization. For scene character recognition, most methods [10,11,15–18] directly extract features from original image and use various classifiers to recognize the character. For scene text recognition, since there are no binarization and segmentation stages, most existing methods [10–13] use various character classifiers to get the candidate character detection results in a multi-scale sliding window fashion. Then various strategies are adopted to get the final word recognition results from the piles of candidate character detections. Wang and Belongie [10] used pictorial structures to spot the final word with the help of a lexicon. Mishra et al. [12,13] built a CRF on the candidate detections to infer the final word via energy maximization. Novikova et al. [14] proposed to use MSER as character candidates and formulate the problem of word recognition as the maximum a posteriori (MAP) inference in a unified probabilistic framework.

2.3. End-to-end scene text recognition

Some methods mentioned above for the individual subtasks have achieved promising performance and quite a few published methods [2,28] for end-to-end text recognition are based on sequential pipeline of the individual steps. However, since each step is independent from each other, there is no possibility to correct errors made by previous stages. Thus, the overall recognition rate of these methods is a product of success rates of each stage. In fact, separating text localization and recognition inevitably leads to loss of information which we could otherwise make use of to further improve the performance.

In recent years, several methods have been proposed to deal with full end-to-end text recognition. Neumann and Matas [29] first detected characters as MSERs and then performed text recognition using the segmentation obtained by the MSER detector. Wang et al. [11] proposed to get the candidate character detections using random ferns in a sliding window fashion and then build the pictorial structures of all the lexicon to find the potential words. This method is rooted in generic object recognition and needs no binarization and segmentation. However, the performance drops significantly as the size of the lexicon increases. Recently, Wang et al. [30] used convolutional neural network (CNN) to train the text detection and