Pattern Recognition Letters 60–61 (2015) 41–47

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis☆

Xiaobing Wang*, Yonghong Song, Yuanlin Zhang, Jingmin Xin

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 16 October 2014
Available online 25 April 2015

Keywords:
Scene text detection
Multi-layer segmentation
Graph cuts
Higher order CRF

ABSTRACT

Text detection in natural scene images is a hot and challenging problem in pattern recognition and computer vision. Considering the complex situations in natural scene images, we propose a robust two-steps method in this paper based on multi-layer segmentation and higher order conditional random field (CRF). Given an input image, the method separates text from its background by using multi-layer segmentation, which decomposes the input image into nine layers. Then, the connected components (CCs) in these different layers are obtained as candidate text. These candidate text CCs are verified by higher order CRF based analysis. Inspired from the multistage information integration mechanism of visual brains, features from three different levels, including separate CCs, CC pairs and CC strings, are integrated by a higher order CRF model to distinguish text from non-text. The remaining CCs are then grouped into words for easy evaluation. Experiments on the ICDAR datasets and street view dataset show that the proposed method achieves the state-of-art in natural scene text detection.

1. Introduction

With the popularity of man-pack imaging devices, such as smart phones, digital images are common in our daily life now. In this tendency, text information extraction from digital images attracts much attention, due to its wide applications in navigation, mobile based text recognition, contented-based image search and so on. A text information extraction system usually consists of two steps [11]: text detection, in which text regions are labeled out, and text recognition, in which text in the labeled regions are retrieved with optical character recognition (OCR) or other technology. For scene text images, a more than 50% improvement of text recognition rate can be achieved by using text detection [8]. Therefore, text detection is important for text information extraction and it is focused on in this paper.

Many methods have been proposed to detect text in natural scene images. Some of them use the features of local image regions (sliding windows) to detect text, called region-based methods [11,17], while some extract text candidates in CC segmentation and identify text in CC analysis, called CC-based methods [6,8,12,18,23–26]. Because CC-based methods have better performance and their detection results can be directly used for text recognition, they attract much attention these years.

CC-based methods consist of two steps: CC segmentation and CC analysis. For CC segmentation, many algorithms have been used, such as Niblack [26], color clustering [12], SWT [6], MSER [24] and so on. Although these algorithms work well in most images, they fail in some complex images. For CC analysis, some text detection methods combined the unary and pairwise features of CCs [18] and some analyzed CCs of different levels such as separate CCs, CC pairs and CC strings in cascade [23,25]. However, integrating all features in different levels are more robust for analysis.

In this paper, a robust CC-based scene text detection method is proposed. When an image is input, a multi-layer segmentation is performed, in which the image is first segmented into super-pixels and then they are labeled into 9 layers to generate candidate text CCs. Non-text of these CCs are removed by a higher order CRF model based analysis and the remaining CCs are detected text. The proposed method is distinguished by the following three contributions:

- A new multi-layer CC segmentation is proposed by integrating the contrasts in different channels.
- The multi-layer CC segmentation is formulated into a multi-labeling problem of super-pixels with a graph cuts based model.
- Inspired from the multistage integration of visual brains, a higher order CRF based CC analysis is used to distinguish text from non-text.

The rest of this paper is organized as follows. A brief overview of the proposed text detection method is presented in Section 2. The details of multi-layer CC segmentation and higher order CRF based

☆ This paper has been recommended for acceptance by Umapada Pal.

* Corresponding author. Tel.: +86 15029752535.
E-mail addresses: wxbxj@stu.xjtu.edu.cn (X. Wang), songyh@mail.xjtu.edu.cn (Y. Song), ylzhangxian@mail.xjtu.edu.cn (Y. Zhang), jxin@mail.xjtu.edu.cn (J. Xin).

http://dx.doi.org/10.1016/j.patrec.2015.04.005
0167-8655/

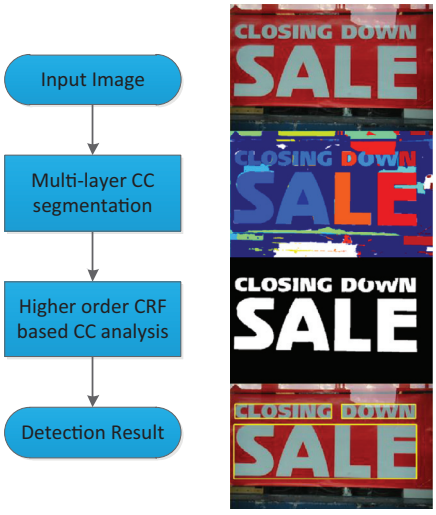

Fig. 1. The flowchart of the proposed text detection method. The CCs generated after multi-layer segmentation are labeled by different colors and the detected words are labeled by yellow bounding rectangles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CC analysis are separately described in Sections 3 and 4. Performance evaluation of the proposed method is shown in Section 5. Finally, conclusions are stated in Section 6.

2. System overview

As shown in Fig. 1, the proposed natural scene text detection method consists of two steps:

1. **Multi-layer CC segmentation**. Given an input image, it is first segmented into super-pixels with the mean shift algorithm. Then, these super-pixels are assigned to 9 labels (from 1 to 9) based on their contrasts, colors and gradients with a graph cuts based model. The layer labeled 9 is the background while the CCs in the other layers are candidate text.
2. **Higher order CRF based CC analysis**. The obtained candidate text CCs are verified to remove non-text ones in this step. Here, a higher order CRF model is employed to integrate the features of separate CCs, CC pairs and CC strings. The remaining CCs are grouped into words based on the distances between them [21].

3. Multi-layer CC segmentation

3.1. Basic idea

The basic idea of the proposed multi-layer CC segmentation is integrating contrasts in different channels to segment an image. Traditionally, segmentation is done several times separately in different channels, such as R channel, G channel and so on, to separate more text from the background. In this study, by integrating contrasts in RGB channels, we show that only one single round segmentation is sufficient. Here the contrast of a pixel x in a channel is computed as follows:

$$Contrast(x) = \frac{I(x) - \mu(x)}{\sigma(x)}, \tag{1}$$

where $I(x)$ is the color value of the pixel x, $\mu(x)$ and $\sigma(x)$ are the mean value and standard deviation of the pixels in the window with width W, as shown in Fig. 2. Therefore, each pixel in each channel can be labeled 1 (brighter) or 0 (darker) according to its contrast. Then, pixels can be labeled from 1 to 8 based on the contrast in RGB channels, as shown in Table 1. In addition, because text usually has certain contrast to the background for easy reading, some non-text pixels with too low contrasts are labeled 9 as the background.

Fig. 2. The local window used for contrast computation. The center of the window is the pixel x.

Table 1
The probable labels of pixels based on RGB contrasts.

Background	R channel	G channel	B channel	Label
Not sure	0	0	0	1
	0	0	1	2
	0	1	0	3
	0	1	1	4
	1	0	0	5
	1	0	1	6
	1	1	0	7
	1	1	1	8
Yes		Any		9

3.2. Implementation

Given an input image, it is first segmented into super-pixels in the multi-layer CC segmentation. Many super-pixel segmentation methods have been proposed, such as mean shift [5], graph-based segmentation [7], SLIC [1] and so on. Because text in scene images usually have different sizes, super-pixels should not be fixed in an approximate size and SLIC is not suitable here. Moreover, considering graph-based segmentation always breaks edges, mean shift is adopted here. With super-pixels as processing objects, the computation is much less. Meanwhile, contrast computed adaptively according to the sizes of super-pixels are more robust than those from a fixed window size.

After super-pixel segmentation, CC segmentation is solved as a multi-labeling problem of super-pixels, which can be formulated by a graph cuts based model [3]. The energy of the labeling $f = \{f_x | x \, is \, superpixel\}$ can be written as:

$$E(f) = \sum_x E_D(f_x) + \sum_{(x,y)} E_S(f_x, f_y). \tag{2}$$

Here $\sum_x E_D(f_x)$ measures the disagreement between f and the observed data, while $\sum_{(x,y)} E_S(f_x, f_y)$ measures the extent to which f is not piecewise smooth.

The data (unary) potentials are computed based on the contrast of these super-pixels. For a super-pixel x, its data potential is $E_D(f_x)$, $f_x \in [1, 9]$ when it is labeled f_x. First, its intensity contrast $T_I(x)$ is computed using Eq. (1). Here $I(x)$ is the mean intensity of pixels in this super-pixel, $\mu(x)$ and $\sigma(x)$ are the mean intensity and intensity standard deviation of the pixels in the window, whose center is the center of x and width is half of the sum of the major axis length and the minor axis length of x. For example, the yellow window in Fig. 3 is computed according to the information of super-pixel "N". Because text usually has certain contrast to the background, super-pixels with too low contrast are background. Here, the contrast threshold T_h is empirically set to 0.4. Based on the idea of integrating the contrast in different channels as shown in Table 1, $E_D(f_x)$ is computed as follows: