Neurocomputing 157 (2015) 153–165

Contents lists available at ScienceDirect

Neurocomputing

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

Text detection approach based on confidence map and context information

Runmin Wang, Nong Sang *, Changxin Gao

Science and Technology on Multi-spectral Information Processing Laboratory, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 12 March 2014
Received in revised form
16 December 2014
Accepted 11 January 2015
Communicated by Xu Zhao
Available online 19 January 2015

Keywords:
Scene text detection
Confidence map
Context information
Texture feature
Connected component analysis

ABSTRACT

Text information plays a significant role in many applications for providing more descriptive and abstract information than other objects. In this paper, an approach based on the confidence map and context information is proposed to robustly detect texts in natural scenes. Most of the conventional methods design sophisticated texture features to describe the text regions, while we focus on building a confidence map model by integrating the seed candidate appearance and the relationships with its adjacent candidates to highlight the texts from the backgrounds, and the candidates with low confidence value will be removed. In order to improve the recall rate, the text context information is adopted to regain the missing text regions. Finally, the text lines are formed and further verified, and the words are obtained by calculating the threshold to separate the intra-word letters from the inter-word letters. Experimental results on the three public benchmark datasets, i.e., ICDAR 2005, ICDAR 2011 and ICDAR 2013, show that the proposed approach has achieved the competitive performances by comparing with the other state-of-the-art methods.

1. Introduction

With the wide use of digital image capture devices, people record their livings more and more conveniently. Correspondingly, the demands based on natural scene images, e.g., content-based image retrieval, visually impaired navigation aid, object recognition and scene understanding, etc., have been grown tremendously in recent years. Texts in nature scenes provide more descriptive and abstract information than other objects. As a result, automatic text detection in natural scenes has been paid wide attentions in recent years. Although the commercial optical character recognition (OCR) systems have achieved great success to separate texts from the documents and understand the contents, text detection in natural scenes is still a challenging problem. The texts in document images are normalized into elegant poses and proper resolutions, while the texts in natural scenes have large variations in text font, size, color and alignment orientation, etc. In addition, text detection in natural scenes is also affected by complex backgrounds, illumination changes, image quality degradation (e.g., some scene texts are shown in Fig. 1).

In the past decade, many researchers have devoted themselves to scene text detection. Numerous methods have been proposed in the literatures, and some survey works can be found in [1,2]. These methods can be roughly divided into two categories: connected component (CC)-based methods and texture-based methods.

CC-based methods based on the facts that the characters in the same text regions exhibit certain properties, such as approximate constant color, proximate pixel value, similar stroke width, etc. Various approaches are used to get the connected components (CCs), e.g., K-means clustering [3], stroke width transform (SWT) [4], stroke feature transform (SFT) [5], extremal regions (ERs) [6], maximally stable extremal regions (MSERs) [7–10], etc. CC-based methods segment an image into a set of CCs, and the final CCs are classified as texts or backgrounds by analyzing their geometrical characteristics. CC-based methods have been proved to be a kind of effective method to detect texts in natural scenes, but these methods meet some challenges when the text is noisy, multi-colored and textured, etc. For the CC-based methods, the core problem is to design a fast and reliable CC analyzer, and some judgment rules based on priori knowledges, e.g., in [4,11,12], aspect ratio, font size, occupation ratio, variance of the stroke width, etc., have been proposed to distinguish the scene texts from the background interferences.

Designing reasonable judgment rules is a very tricky thing for the CC-based methods. The strict judgment rules can improve the

* Corresponding author.
E-mail addresses: runminwang@hust.edu.cn (R. Wang), nsang@hust.edu.cn (N. Sang), cgao@hust.edu.cn (C. Gao).

http://dx.doi.org/10.1016/j.neucom.2015.01.023
0925-2312/

Fig. 1. Some texts in natural scenes.

precision rate of text detection, but it often leads to a drop in the recall rate, and vice versa. Throughout the existing CC-based methods, the judgment rules are built just based on the CCs' geometrical characteristics, which makes the methods lack robustness. In fact, the adjacent CCs in the same text line usually have similar height, approximate constant color, proximate pixel value, similar stroke width, etc., these contexts between the adjacent CCs are valuable information for us. Differently, most of the conventional CC-based methods design the judgement rules only based on the candidate's geometrical characteristics, while some loose judgment rules and the confidence map are integrated to remove the background interferences in our work. Precisely speaking, the loose judgment rules are adopted to preliminarily remove the candidates that are most unlikely to be texts, and then, the confidence map is adopted to eliminate the tricky interferences in further. In our work, the judgment rules are built on the candidates' geometrical characteristics, and the confidence map model proposed in our work are composed of the similarities between the adjacent CCs and the text likelihood of the seed candidate evaluated by the texture-based method.

Texture-based methods assume that text regions have distinct textural features compared with the backgrounds, and these methods are efficient in dealing with the text detection problem in complex backgrounds. The texture-based methods generally consist of two steps: the features extracted from each candidate region are fed into the trained classifier at the first step, and then the text likelihood of the candidate region is estimated by support vector machine (SVM) [13], AdaBoost [14,15], and artificial neural networks (ANN) [16], etc. In the texture-based methods, the feature design is very important. There are various kinds of feature vectors, e.g., gradient edge features [17], T-HOG [18], Gabor features [19], etc., have been designed to describe the candidate regions.

Unlike face, which is normally composed of eyes, nose, cheeks and chin, text region consists of different words with a set of character combinations. Thus it is difficult to design a universal text descriptor to consider text regions as a general class by contrast with the backgrounds. To solve this problem, two kinds of ways are mainly adopted in our work. One way is the confidence map, which has been mentioned in the above paragraphs, and the other is the convolutional neural networks (CNN) classifier proposed in [20] based on the advantage that the classifier is trained by the features independently chosen by an unsupervised algorithm.

There are a few text regions may be lost after character candidates verification processing. The missing text regions will reduce the recall rate, and they will also affect subsequent text grouping and word partition processing. To solve this problem, a heuristic rule based on the context information is designed in our work to regain the missing text regions. In addition, it should be pointed out that a concise method is proposed to form the text lines in our work. For each CC, the morphological structuring element is dynamically selected based on the CC's height, and then the morphological closing processing is performed between the CC and its adjacent CCs. Finally, these subimages are merged to obtain the text lines.

The core task of text detection is to find the text location information in the image. The detection performance and execution efficiency are the two important indexes to evaluate the algorithm. The CC-based method is relatively fast, and the texture-based method using trained classifiers has good performance in dealing with complex backgrounds. In this paper, the texture-based method and the CC-based method are reasonably integrated by using the context information to robustly detect the texts with various of scales, colors, and clutter backgrounds from natural scene images. Our main contributions of this work include the following aspects: