An End-to-End scene text recognition system and its application in web security system

【Abstract】

The number of images is booming. Many application systems need to extract the text contents in web images in order to do image-content based analysis. For example image search engines, street-view based map guidance APPs. And even web security systems, which is proposed in this paper. For these systems, extracting texts from images is a core components. //pre-required.

Extracting texts from Natural images is a hard job, It usually involves in text detection, localization and recognition. For text localization, among which text Localization is a very important step which highly affect the performance of OCR , especially when handling with natural scene images. Recently MSER region detector has shown promising efficiency in scene text detection.

In this paper we proposed a robust end-to-end text exaction system . First an improved MSER method is proposed, the detected MSER regions are processed by a text candidate region filter made up with CNN classifier and a character recognition feedback component. Second, a region position and context based clustering algorithm is used to merge the text regions into words and text-lines. Finally they are recognized using a CNN classifier. The experiments show our algorithm has high efficiency in speed and state-of-the-art accuracy both in localizing and recognition. The comparison of this method with former methods on the ICADR 2011 Dataset is also demonstrated.

And the end-to-end extraction system has been used in a web-security system.

【Keywords】 Scene Text Localizing , MSER , CNN ,

* 1. 1 Introduction

The number of images is booming. Many application systems need to extract the text contents in web images in order to do image-content based analysis. For example image search engines, street-view based map guidance APPs. And even web security systems, which is proposed in this paper. For these systems, extracting texts from images is a core components.

Extracting texts from Natural images is a hard job. As Fig.1 shows, an image text exaction algorithm usually involves in text detection, localization and recognition , among which text Localization is a important pre-processing step which highly affect the performance of OCR , especially when handling with natural scene images with backgrounds unstrained and texts in varied styles.While character recognition has become a more and more resolvable problem, an example is that even the CAPTCHA which is designed to prevent machine from recognition is threatened by deciphering , text detection and localization is still a problem which highly effects the quality of the finally retrieved text information from images.

Input images

Text detection and localization

Recognition(OCR)

OCR Results

1.1 Architecture of an text exaction system

Currently, The existing methods of text localization and extraction algorithms can be roughly categorized into two groups: sliding-window based and connected component (CC) based[[[1]](#endnote-2)]. Sliding-window methods are often slow for it has to detect the images in different scales.

Connected component based methods [[[2]](#endnote-3), [[3]](#endnote-4), [[4]](#endnote-5), [[5]](#endnote-6)] typically detect CCs from the image with approaches like edge detection, color clustering, or stroke width transform, then group the pixels into regions using text candidate detectors, classify out non-text regions using typically classifiers, and group extracted regions together into words or sentences. Its grouping strategy is based on the suppose that pixels belonging to the same character have some similar properties, for example color, gradient, stroke width and so on. Connected component based methods can do word or character

Recently, Many papers [[[6]](#endnote-7)12] use Maximally Stable Extreme Regions (MSERs) [[[7]](#endnote-8)] based methods , some of them have reported good performance on the ICDAR 2011 Robust Reading Competition Dataset [[[8]](#endnote-9)]. An ER is a extremal region which has extremal property of intensity than its outer boundary. And an ER of threshold θ(R θ) is such an region: ∀p ∈ R, q ∈ ∂R : C(q) > θ ≥C(p). An MSER is an particular case of an Extremal Region whose size remains virtually unchanged over a range of thresholds. MSER is very effective in text detection since text regions usually have a higher contrast in the image in order to be easily read by humans [6].

But MSER also have problems. First a image can contains a lot of high contrast areas, therefore it detects large amount of false positive text regions which need to be eliminated on the next. Second, as contrast based, it have a high miss rate on low-contrast images and the detection. Third, like other connected component based methods, the detected regions need to be clustered into words and text-lines as a final result of text recognition.

[] proposed a real-time scene image text localization algorithm based on ER classification. Their method first extract ER(extremal regions) and ER descriptors in a fast and incremental computed way , then the extracted ERs' several descriptors are served as features for a two-stage classification to detect text candidate regions. Their experiments show their proposed method have problems when handling with: (a) Characters with no contrast. (b) Multiple characters joined together. (c) A single letter (the claim that the McDonald’s logo is a letter “M” as defined by the annotation is questionable), the recall rate is low.

[] proposed a method which do “Perceptual Organization Clustering” for MSERs. They argued that text extraction problem could be posed as the detection of meaningful groups of regions In details the extracted MSER regions are grouped together in a bottom-up manner guided by similarity evidence obtained over various modalities such as color, size, or stroke width among others.The result show it has a state-of-the-art performance at text localization, especially when handling with images containing Chinese texts. The disadvantage is that different words are easily clustered together into one segmentation, and the precision is improvable relative to its recall. And the words are clustered based on their styles(such as stroke width, hole area rate) then on their position property.

In this paper we proposed a scene text localization algorithm which has three points : first novel features are proposed which will be used by CNN for classification and text regions are filtered out; second the regions will be quickly recognized by a recognizer using template matching and the recognition result will be sent back to filter out non-text regions. Third the region position and the character context will be used to cluster the regions into words and text-lines.

The following sectors are organized as follows: In sector II we will describe our improved MSER based text localization algorithm. In sector III we will make experiments to demonstrate its efficiency and accuracy in processing scene images. We will also make comparisons between our algorithm and the former algorithms. In sector IV we will make a summary of our work.

* 1. 2 algorithm

Our algorithm can be organized as the flow-chat in figure 2.

* + 1. A. Tuned MSER text detection

MSER detection algorithm is widely used in object detection. The core of the method is the contrast of objects from the background. But since many regions have a high contrast and this method result in many non-text regions. As proposed by [[[9]](#endnote-10)],it is possible to exploit the inclusion relation between ERs to incrementally calculate ER, and in this way a Linear Time Maximally Stable Extremal Regions algorithm is developed.

MSER detector has some disadvantages especially when handling with intensity varied regions, as figure 1 shows, the MSER doesn't work well when the character 'O' and 'W''s intensity close to the are recognized by human because they have a different color from

figure 1: ＭSER detector has missed "O" and "W" from "DOWN" as a result of character's intensity change

there are many problems with MSER, first its recall face problems when handling with special pictures.

compute descriptor values.

* + 1. B. text region filtering

With the exaction of tuned MSER detector, a large number of CC regions are obtained. To reduce the calculation complexity, we first do a simple filtering to eliminate non-text regions based on several rules which check regions' geometric features. The features include : 1. commonly used aspect radios (0.1~1); 2. area .regions whose area is larger than half of the whole image or smaller than 4px\*4px will be eliminated ; 3. hole numbers. Typically, English characters have no more than two holes, here we use three as a threshold. 4. corner number. For an English character, corner number is between 4 and 12. With this rule, we can tell natural objects from human made characters.

To improve the precision of the text localization, non-text regions need to be removed. In this paper, we realize high precision with three points:

1 filters with several rules

As we have phenomol that isolate regions are often meaningless, so we did a text-line clustering which cluster the regions based on their positions. That is, text regions that have a neighbor would be more likely reserved while separated regions eliminated. on one line will be more likely clustered together. And absolute regions are filtered out. Experiments show it has a good performance.

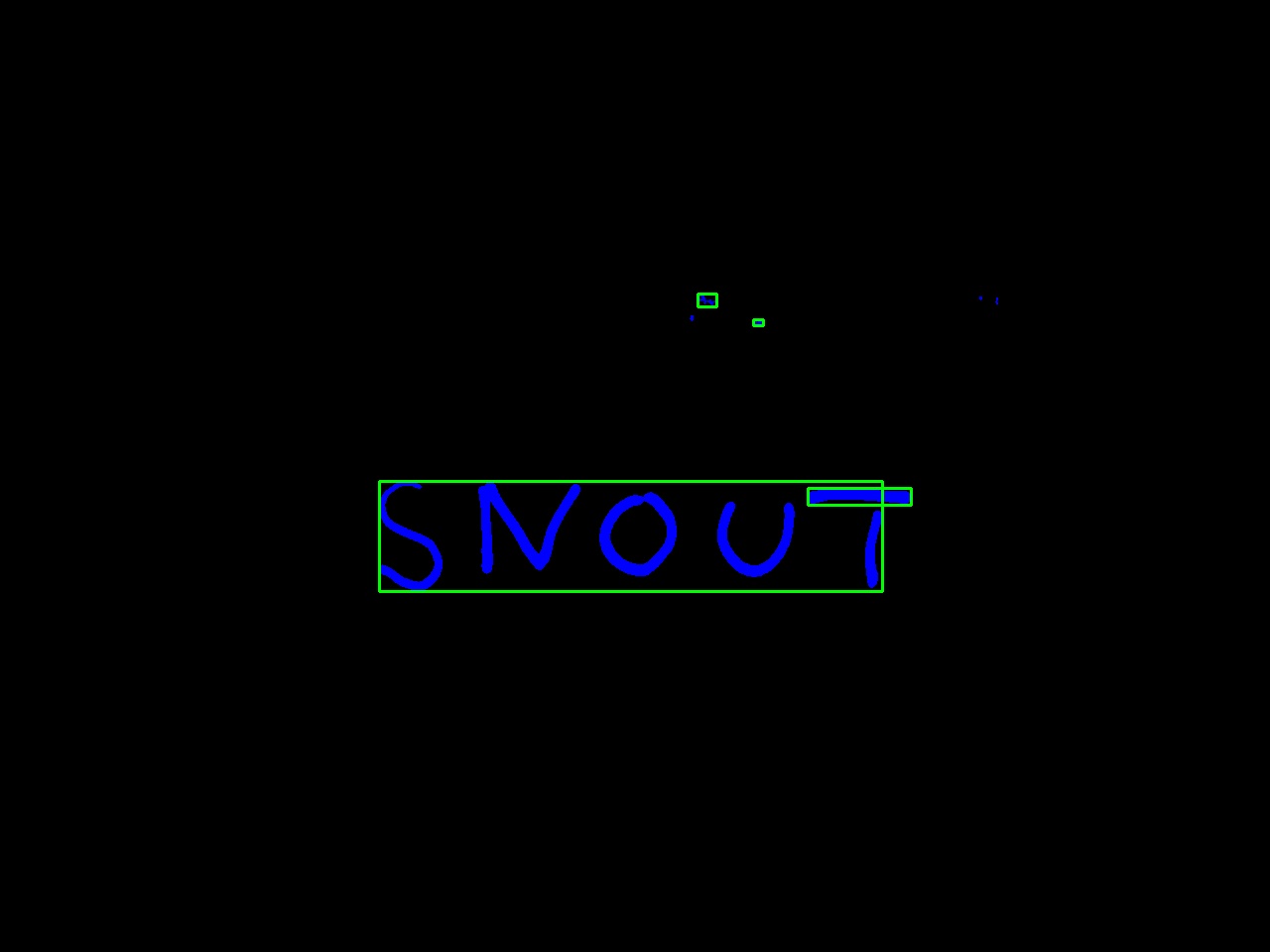


fig. 1: isolate regions are filtered out while regions in text-line are reserved

**C Text region classification**

To eliminate text regions before further recognition , simple filtering is not enough. Many MSER based papers do region classification, to classify out non-text regions based on several text region features. The classifier can also merge the regions into words and lines by the way.

D Character recognition feedback

In this paper, motivated by [] which said recognition results should provide valuable feedback for text detection, we involve a character region recognition algorithm whose output will become the evidence for the region as a text-region.

Our character recognition algorithm uses a CNN classifier to classify the character regions.

* 1. 4 Experiment Results

To evaluate our text detection algorithm,

our algorithm is both test on the ICDAR 2003 and ICDAR 2011 Dataset. On the ICDAR 2011 Dataset, the

* + 1. 4.1 ICDAR Text Detection

On ICDAR 2003, We show the text detection performance on the dataset in Table 1.

1. X.-C. Yin, X.-W. Yin, K.-Z. Huang, and H.-W. Hao,“Robust Text Detection in Natural Scene Images,”IEEE Trans. Pattern Analysis And Machine Intelligence,vol. , no. , [↑](#endnote-ref-2)
2. B. Epshtein, E. Ofek, and Y.Wexler. Detecting text in natural scenes with stroke width transform. In CVPR 2010, pages 2963 –2970. [↑](#endnote-ref-3)
3. L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In ACCV 2010, volume IV of LNCS 6495, pages 2067–2078, November 2010. [↑](#endnote-ref-4)
4. L. Neumann and J. Matas. Text localization in real-world images using efficiently pruned exhaustive search. In ICDAR 2011, pages 687–691, 2011. [↑](#endnote-ref-5)
5. Y.-F. Pan, X. Hou, and C.-L. Liu. Text localization in natural scene images based on conditional random field. In ICDAR 2009, pages 6–10. IEEE Computer Society, 2009. [↑](#endnote-ref-6)
6. 2unzhao Shi, Chunheng Wang ,Baihua Xiao, Yang Zhang, Song Gao, “Scene text detection using graph model built upon maximally stable extremal regions”, In Pattern Recognition Letters 34 , 2013 [↑](#endnote-ref-7)
7. J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” in Proc. British Machine Vision Conference, vol. 1, 2002, pp. 384–393. [↑](#endnote-ref-8)
8. A. Shahab, F. Shafait, and A. Dengel, “ICDAR 2011 robust reading competition challenge 2: Reading text in scene images,” in Proc. Int. Conf. Document Analysis and Recognition, 2011, pp. 1491–1496. [↑](#endnote-ref-9)
9. Linear Time Maximally Stable Extremal Regions [↑](#endnote-ref-10)