PROCEEDINGS OF SCIENCE

Text Detection in Natural Scenes Based on Maximally Stable External Region and Deep Convolutional Network

Tianhao Li[1][2]
Department of Information Science and Electronic Engineering, Zhejiang University, 310027 Hangzhou, China
E-mail: wait4pumpkin@gmail.com

Huimin Yu[3]
Department of Information Science and Electronic Engineering, Zhejiang University, 310027 Hangzhou, China
E-mail: yhm2005@zju.edu.cn

In this paper, an effective scene text detection algorithm is proposed based on Maximally Stable External Region (MSER) and deep convolutional network. This algorithm is competitive with the best existing scheme widely used on benchmark International Conference on Document Analysis and Recognition (ICDAR) 2013. The main contribution of the paper lies in two aspects. Firstly, an efficient text/background classifier is designed on the basis of convolutional neural network, which outperforms the artificial design features. What's more, synthesis data is used for constructing the training set so as to avoid over-fitting; secondly, an analysis method based on MSER is presented to drop out the majority background patches and extract valid character areas. On the basis of analysing connected components, the size of candidate character is determined instead of the time-consuming scale scanning.

CENet2015
12-13 September 2015
Shanghai, China

[1]Speaker
[2]This work was supported by Zhejiang Province Science and Technology Plan Project (Program No.: 2013C310035) and NSFC under Grant No.: 61471321 and National Key Basic Research Project of China (973 Program No.: 2012CB316400).
[3]Corresponding Author

PoS(CENet2015)085

1.Introduction

The scene text extraction plays an important role in the image understanding because of its great semantic value. The filming location could be inferred from recognizing the scene text, such as road signs and signboards, which contributes to the video surveillance and the location-based services. On the basis of numerous researches carried out in recent years, the problem of scene text detection is far from solved due to its complexity. The difficulty lies in several aspects, including the sophisticated of scene and the stochastic position of text. In addition, the scene text is varied in size, font style and color; furthermore, the rotation, perspective transform also makes the problem challenging due to various shooting angle.

In this paper, an effective algorithm for the scene text detection is proposed with its contributions listed as follows.

(1)Inspired by the excellent performance of CNN (Convolutional Neural Network) on ImageNet image recognition problem [1], an effective Text/Background Classifier is proposed. Several optimized measures are imposed for the text detection problem.

(2)In order to decrease the time complexity of the algorithm, especially the computation of convolution classifier, a method for extracting character candidates is proposed based on MSER.

(3)An effective method for synthesizing labeled samples is presented so as to avoid over-fitting for deep neural network with large number of parameters.

Related Work

Methods of the scene text detection can be generally divided into two categories, the connected components based method and the texture based method. The main idea of the connected components based method is to construct text regions by grouping valid connected components by means of analysis or classification. The greatest advantage of connected components based methods is to calculate the efficiency; and through region extraction, the time-consuming scale scanning procedure is avoided, for example, Neumann proposed a cascade classifier to determine if an External Regions (ERs) belongs to text [2], and Chen employed the edge-enhanced maximally stable extremal regions for text detection [3]. Li used the width transform to determine valid text areas [4]. Recently, Shi used the graph model built upon maximally stable extremal regions to classify the text region [5]. As to the texture based method, a sliding window scans through the whole image, then each patch within the window is classified independently. Compared with connected components based method, the texture based method features superiority in terms of feature expression but at the expense of computational efficiency. There's a typical example, which applied Adaboost with several features including derivatives, histogram, etc. to classify whether a patch is a valid text region [6]. Coates achieved better result based on unsupervised feature learning [7].

2. Text Candidates Extraction Based on MSER

MSER (Maximally Stable External Region) is a widely used region descriptor, which is designed on connected components under different thresholds and only keeps "stable" regions. The stable constrain, briefly means the area change is relatively small under adjacent thresholds. A MSER tree, as shown in the middle of Fig. 1, will be generated after MSER extraction; however, the constraint of MSER is too weak for the text detection problem, a number of background regions are included in MSERs. In addition, one character may contain several MSERs under different thresholds, as shown in Fig. 1. Thus, the valid character region in MSER tree must be determined to construct the text candidate regions, and background regions should be removed as much as possible.

Figure1: Example of MSER Extraction.

Left: original image; middle: the extracted MSER tree; right: valid character region.

In order to extract valid text regions, an extraction algorithm based on probability threshold is proposed. The probability is calculated by a linear classifier, which is defined as below,

$$prob = \sigma(\boldsymbol{W}^T \boldsymbol{x} + b) \tag{2.1}$$

where $\boldsymbol{W}$ and b refer to the weight and bias of the classifier respectively; σ is the activation function, and sigmoid function is used here; $\boldsymbol{x}$ is the feature vector extracted from connected component defined as below:

- Ratio between width and height of bounding box.
- Ratio between contours perimeter and contours area.
- Standard deviation of RGB value within the region (values are scaled to $[0,1]$).
- Standard deviation of stroke width [4].
- Area ratio between contours and convex hull.
- Area ratio between inner contours and outer contours.
- The mean depth of convex defect.

Fig. 2 shows the complete extraction algorithm. There are two cases. When no child belongs to the processing node T, the node will be added if its probability exceeds specific threshold τ; otherwise, all the children will be processed iteratively and combined with a set named V. If one of the probability in set V exceeds the probability of node T, the set V will be returned, or T will be returned if not. The threshold τ is set up during the classifier training.

The proposed algorithm aims at solving the two problems arising from the beginning of this section. A set of feature is designed to strengthen the constraint of MSER, including the commonly connected components analysis and the text specific attributes, such as stroke width and color. Thanks to the selected features and probability threshold, the majority background areas are eliminated. In addition, valid character regions must be extracted from the tree structure because MSER is a hierarchical descriptor. As shown in Fig. 1, the region "CA" and its children, "C" and "A", are classified as valid regions while the former is not a valid character. A selection criterion based on parent-children comparison is set up in order to decide which one is more likely to be valid character. The principle is simple. If the max of the children's probabilities exceeds the parent's, the parent is ignored and all children are chosen; otherwise, the parent is chosen and all children are ignored if not. Based on the selection principle, valid character regions are extracted from hierarchical MSER tree.

```
function extract(T):
    V = {}
    if T.numChildren > 0 then
        for child in T.children loop
            V += extract(child)
        end
        if max-prob(V) < T.prob then
            if T.prob > τ then
                V = {T}
            else
                V = {}
            end
        end
    else if T.prob > τ
        V = {T}
    end
    return V
end
```

Figure 2 : Proposed Algorithm for Extracting Valid Character Regions from MSER Tree.

3. Character Classifier Using Deep Convolutional Network

The feature extraction is the key to general image recognition tasks and it determines the performance of the algorithm that can be reached. Although some features, such as SIFT and HOG, show good results in several experiments, the generalization is not promised for artificial design features. More importantly, these features are designed manually heavily depending on experience. Deep convolutional network tries to learn the best classification features provided the training set, which outperforms artificial design features in several benchmarks [1].

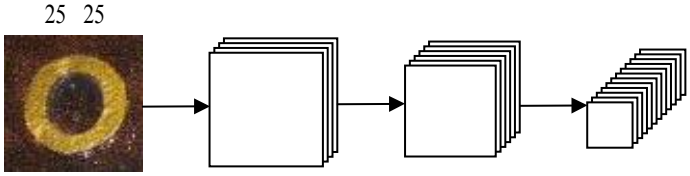

Figure 3: Proposed Architecture of Deep Convolutional Network.

Fig. 3 shows the proposed architecture of convolutional neutral network. The network consists of three convolutional layers. A convolutional layer contains K filters, which maps input $\boldsymbol{x} \in \mathbb{R}^{m \times n \times L}$ to a group of features, $\boldsymbol{h} \in \mathbb{R}^{p \times q \times K}$ where each filter processes as follows:

$$\boldsymbol{h}^k{}_{\boldsymbol{W},b}(\boldsymbol{x}) = \sigma\Big(\sum_{l}^{L} \boldsymbol{W}^k * \boldsymbol{x}^l + b^k\Big) \tag{3.1}$$

$*$ is the convolution operator; $\boldsymbol{W}$ and b are filter template and bias respectively; σ is nonlinear activation function, which maxout [8] with group of 2 is used here as shown in Fig. 4. All filters are in size $9 \quad 9$, and number of filters in each layers are 32, 48 and 96 respectively as shown in Fig. 3.

Compared with typical convolutional neural network [1], several improvements have been made as to text detection problem. Firstly, the pooling layer is removed as to small patch size ($25 \quad 25$) so as to achieve calculation efficiency; secondly, the maxout is used for activation function instead of normal nonlinear function, like sigmoid and ReLU, since the maxout shows outstanding result in several benchmarks [8]; moreover, the selection of max activation in maxout helps avoid over-fitting when the polling layer is eliminated.

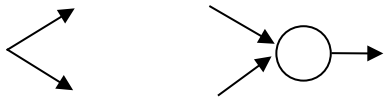

Figure 4: Procedure of Maxout Activation with Group of 2.

Training Set: in order to avoid over-fitting, the training set is constructed by synthesis samples because the labeled samples are rare. The generation produce is shown as follows. A character with random font style is aligned at the center, and random deformation is processed including rotation, translation and perspective transformation. Then an image patch sampled from background dataset is synthesized with the character image. To make the training sample more challenging, the background patches whose mean of grey level is bigger than 0.8, or standard deviation is lower than 0.2, are ignored. Fig. 5 shows part of synthesis samples.

Figure 5 Part of Synthesis Samples.

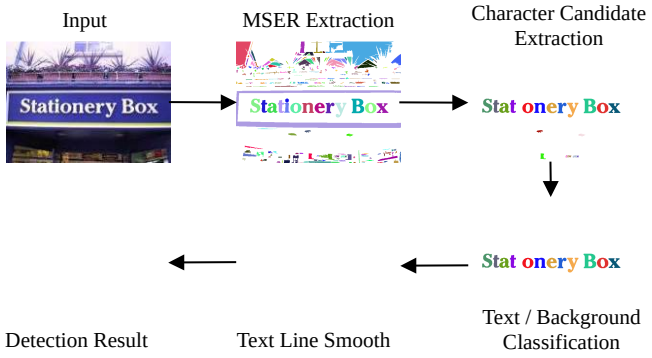

Fig. 6 Processing Pipeline for Proposed Glgorithm.

4. End-to-End Pipeline

As to the conventional object detection tasks, the sliding window method is usually used. Every image patch within the window is extracted and feeds to the classifier independently, then the classifier determines if the input belongs to the target object. The idea of sliding window method is simple; however, the calculation is not efficient enough, especially when the convolution neural network is adopted as the classifier because the typical serial computing model is not good at convolution operation. On the one hand, in order to locate the target precisely, each sliding window should have some overlap; on the other hand, the same detection procedure needs to be processed independently under various scales of original image so as to detect targets in different sizes. Both of the operations will complicate the calculation.

Two methods can be used to reduce computational complexity of detection procedure. Firstly, smaller patch size would be helpful, as described in Section 3. Secondly, reduce the amount of patches as far as possible; therefore, MSER extraction and connected component analysis are applied before precise detection by convolutional neutral network in order to eliminate most of background patches. In addition, on the basis of character candidate extraction, the scale of target would be determined; this it is not necessary to scan at different dimensions in post processing.

Fig. 6 shows the processing flow of the proposed algorithm.

MSER Extraction: extract all areas satisfied MSER constraint and ignore other regions for the next processing[9].

Character Candidate Extraction: valid regions are extracted from MSER tree through probability threshold (Section 2). All candidate regions are then scaled to same size. The rescaling approach ensures that no scanning is needed for detecting the text areas of different dimensions.

Text/Background Classification: each candidate is classified if it belongs to text areas by the pertained convolutional neutral network (Section 3).

Text Line Smooth: run length smoothing algorithm is processed to cluster character regions into strings: adjacent regions are connected if the space between them is less than $2\mu + 0.5\sigma$, where μ and σ refer to the mean and standard deviation for centers of character regions in one row accordingly. As some characters with simple shapes, like "i" and "l", are easily classified as noise in preceding processing, the smoothing approach helps join the misclassified characters into complete word.

Detection Result: extract connected components from the smoothing result.

5. Experiments

The training set for the linear classifier as described in Section 2 consist of 4,786 character patches extracted from ICDAR 2013 training set and 5,000 background patches extracted from VOC 2012 [10-11]. The threshold τ is set to 0.15 for high recall (98.1%) at the expense of low precision (73.0%), as shown in Fig. 7.

The deep convolutional network is implemented based on Caffe [12], a deep learning framework. 1,000 samples for each class, including digits, upper and lower case characters, 62,000 in total, are synthesized as described in Section 3. 65,000 negative patches are randomly selected from VOC 2012 and are scaled to the same size. Fig. 7 shows the precision/recall curves of deep convolutional network, 87.5%/91.3%/89.4% (precision/recall/F1-measure) as to the synthesis data, 75.1%/85.5%/80.0% as to ICDAR 2013 training set.

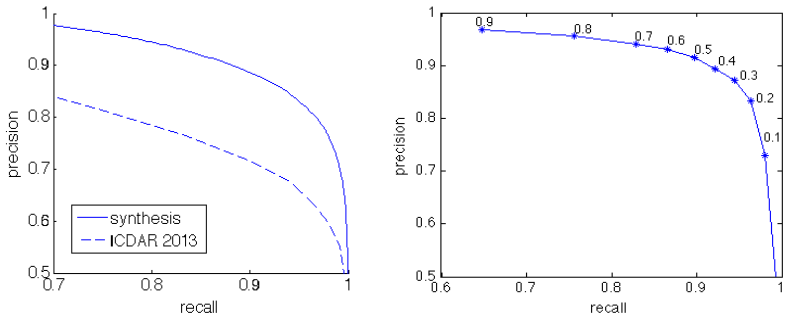

Figure 7: Precision/Recall Curves.

Left: precision / recall curves of deep convolutional network. Right: precision / recall curve of linear classifier with different threshold.

Fig.ure 8: Some of Correct Detected Samples.

Method	Precision	Recall	F1-measure
USTB_TexStar	88.47	66.45	75.90
Proposed	90.10	65.01	75.53
TextSpotter	87.51	64.84	74.49.
CASIA_NLPR	78.86	68.24	73.18

Table 1: Text Detection Precision (%), Recall (%) and F1-measure (%) for Different Algorithms.

As to the end-to-end processing, the same criterion as ICDAR 2013 is used, including precision, recall and F1-measure. As shown in Table 1, the proposed algorithm is competitive with the winner of ICDAR 2013 competition [10]. As the experimental result (Fig. 8) shows, the proposed algorithm highlights sound adaptability for different kinds of scene text, even with rotation, deformation and distortion; however, the detecting text in low contrast with complex background is still challenging.

6. Conclusion

In this paper, an effective scene text detection algorithm is proposed based on MSER and deep convolutional network. An analysis method based on MSER is presented, drops out the majority background patches and extracts all valid character areas. All candidate text regions are then classified by efficient text/background classifier based on the convolutional neural network. Experimental results show that the proposed algorithm is competitive with the best existing schemes on widely used benchmark ICDAR 2013.

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton et al (2012). *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems. MIT Press, Cambridge, MA, USA, 2012: 1097-1105.

[2] L. Neumann, J. Matas (2012). *Real-time scene text localization and recognition. Computer Vision and Pattern Recognition (CVPR)*, 2012 IEEE Conference on. IEEE Computer Society, Los Alamitos, CA, USA, 2012: 3538-3545.

[3] H. Chen, S.S. Tsai, G. Schroth D.M. Chen, R. Grzeszczuk, B. Girod (2011). *Robust text detection in natural images with edge-enhanced maximally stable extremal regions.* Image Processing (ICIP), 18th IEEE International Conference on. IEEE, Piscataway, NJ, USA, 2011: 2609-2612.

[4] Y. Li, H. Lu. *Scene text detection via stroke width. Pattern Recognition (ICPR)*, 21st International Conference on. IEEE, Piscataway, NJ, USA, 2012: 681-684.

[5] C. Shi, C. Wang, B. Xiao Y. Zhang, S. Gao (2013). *Scene text detection using graph model built upon maximally stable extremal regions.* Pattern Recognition Letters 2013, 34(2): 107-116.

[6] J.J. Lee, P.H. Lee, S.W. Lee A. Yuille, C. Koch (2011). *AdaBoost for Text Detection in Natural Scene. Document Analysis and Recognition (ICDAR),* 2011 International Conference on. IEEE Computer Society, Los Alamitos, CA, USA, 2011: 429-434.

[7] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, W. Tao et al. *Text detection and character recognition in scene images with unsupervised feature learning.* Document Analysis and Recognition (ICDAR), International Conference on. IEEE Computer Society, Los Alamitos, CA, USA, 2011: 440-445.

[8] I.J. Goodfellow, D. Warde-Farley, M. Mirza A. Courville, Y. Bengio (2013). *Maxout networks.* Proceedings of the 30th International Conference on Machine Learning (ICML-13). ACM, New York, NY, USA, 2013: 1319-1327.

[9] D. Nistér, H. Stewénius (2008). *Linear time maximally stable external regions.* Computer Vision–ECCV, Springer, Berlin, German, 2008: 183-196.

[10] D. Karatzas, F. Shafait, S. Uchida M. Iwamura, L. Gomez i Bigorda, S. Robles Mestre et al (2013). *ICDAR 2013 robust reading competition. Document Analysis and Recognition (ICDAR),* 2013 12th International Conference on. IEEE, Piscataway, NJ, USA, 2013: 1484-1493.

[11] M. Everingham, L. Van~Gool, C.K.I. Williams, J. Winn, A. Zisserman(2012). *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.* URL http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[12] Y. Jia, E. Shelhamer, J. Donahue S. Karayev, J. Long, R. Girshick et al (2014) *Caffe: Convolutional architecture for fast feature embedding.* Proceedings of the ACM International Conference on Multimedia. ACM, New York, NY, USA, 2014: 675-678.