Advances in Computer Science Research, volume 65

3rd International Conference on Communications, Information Management and Network Security (CIMNS 2018)

Natural Scene Text Detection Based on MSER

Kun Wang[1], Guokuan Li[1, 3, *], Xujun Liu[1], Jingkun Yan[1], Shuli Li[1] and Hao Huang[2, 3]

[1]Wuhan National Research Center for Optoelectronics, Huazhong University of Science and Technology, China

[2]School of Software Engineering, Huazhong University of Science and Technology, China

[3]Shenzhen Research Institute, Huazhong University of Science & Technology, Shenzhen, China

[*]Corresponding author

***Abstract*—Scene text detection has important applications in the fields of intelligent transportation, industrial automation, multimedia retrieval and so on. This paper employs the improved MSER algorithm combined with convolutional neural network for scene text detection. Gradient amplitude enhancement processing is used to enhance the text boundary before a combinational suppression strategy is applied to filter out coincident regions, approximately coincident regions and nested regions. Then the Char-CNN classifier is designed to classify the candidate regions. A hierarchical clustering algorithm is used to merge the candidate regions, and finally generate the text position information. We evaluate the algorithm performance on the ICDAR2013 dataset. The results show that the improved MSER algorithm increases the recall rate of the character region from 88.9% to 90.2%, and the proportion of character regions in the candidate regions increases from 3.25% to 35.19%. And the classification accuracy of Char-CNN is 93.6%. The recall and accuracy rate of the algorithm are 0.68 and 0.85 respectively, and the F-Measure value is 0.76. Compared with existing scene text detection algorithms, the proposed algorithm has a competitive overall performance.**

Keywords—text detection; maximally stable extremal regions; convolution neural network; hierarchical clustering

I. INTRODUCTION

Natural scene text detection refers to detecting and locating text regions in natural scene images. The text information in the scene image is an important semantic information, which can briefly express the main content of the image. Scene text detection has important significance for image understanding, analysis and retrieval. The practical applications of scene text detection is mainly found in the fields of image retrieval, intelligent transportation, commercial paper processing, image encryption, automatic translation, and blind navigation etc.

The methods for classifying the scenes can be classified into two categories: methods [1-4] based on a sliding window and methods [4-8] based on a connected region. For the former category, the input image is first scaled on different levels, and then window sliding of different sizes is used on each scale image to extract pre-designed edges, textures and other local structural features for each sliding window. Based on these feature values, a feature vector is formed for training classifier to determine whether the sliding window is a character region. The latter category is based on the consistency of the character region in terms of brightness, color, stroke width characteristics etc. First, the candidate region detection algorithm detects a large number of connected regions of the suspected characters, and then the manually designed features such as edges, textures and geometric strokes of characters are used to classify the candidate character regions, filter out the character connected regions, and finally merge the character regions into text lines by combining the spatial position and color relationship of the character regions, thereby completing the text positioning.

Based on the existing scene text detection algorithm, this paper combines the improved MSER algorithm and convolutional neural network to detect the scene text. The work done in this paper mainly includes (1) optimizing the classical MSER algorithm, and using the gradient amplitude enhancement processing to enhance the blurred text boundary of the scene image. The coordinate information and geometric information of the candidate regions are utilized by a combined suppression strategy for filtering the coincident regions, the approximate coincident regions and the nested regions. (2) Design of a convolutional neural network Char-CNN, which is specifically used for candidates region classification; (3) Use of a cohesive hierarchical clustering algorithm to generate the final text.

II. NATURAL SCENE TEXT DETECTION METHOD

The flowchart of MSER-based scene text detection algorithm proposed is shown in Figure I, which mainly includes three main steps: candidate region extraction, candidate region classification and character region merging into text lines.

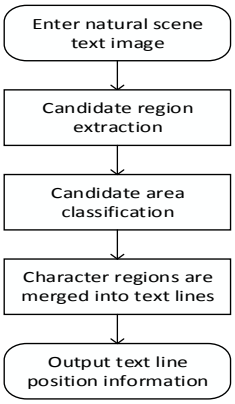

FIGURE I. FLOWCHART OF PROPOSED ALGORITHM

The candidate region extraction process is to extract the connected regions of the possible character regions. In this paper, an improved MSER algorithm is designed to extract candidate regions. Firstly, the gradient of the scene image is used to improve the contrast of the text boundary, and then the

coordinates of the candidate regions are used to design a combined suppression strategy for filtering coincident regions, approximate coincident regions, and nested regions. The method can filter out a large number of non-character regions in the candidate region extraction phase. This paper designs a Char-CNN binary classifier based on convolutional neural network to classify the candidate character regions and uses semi-automatic method to establish a dataset containing 1600 positive samples and 8000 negative samples to train Char-CNN. The Char-CNN network designed in this paper can be very effective. Finally, the character regions are merged to generate the text. The hierarchical clustering algorithm is used according to the position of the character region centroid, and distances between classes and the distance within the classes.

A. Improved Scene Text Candidate Region Extraction Algorithm

The flowchart of improved MSER algorithm proposed in this paper is shown in Figure II. The traditional MSER algorithm is improved from two aspects. On the one hand, the gradient amplitude enhancement is applied to the grayscale of the original scene image, which improves the contrast of blurred boundaries. On the other hand, a combined suppression strategy is applied to filter the coincident region, the approximate coincident region and the nested region to improve the proportion of the positive candidate regions.

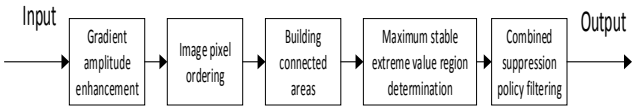

FIGURE II. FLOWCHART OF IMPROVED MSER ALGORITHM

1) Gradient amplitude enhancement improves scene text boundaries

a) Converting an input image from an RGB color space to a grayscale image

b) using the Sobel operator to obtain a gradient amplitude map G;

c) Combining the gradient amplitude map G to enhance the original grayscale image Gray. Finally, Gray1 is normalized to the range [0,255];

d) Area detection using the MSER algorithm.

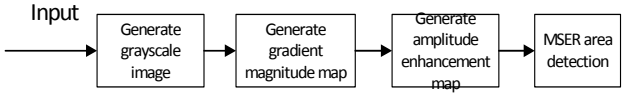

FIGURE III. GRADIENT AMPLITUDE ENHANCEMENT

2) Candidate region filtering based on combined suppression strategy

First, define the coincident candidate region, the approximated coincident candidate region, and the nested candidate region. The candidate region obtained by the improved MSER algorithm is defined as a sequence $Q_1, Q_2, ..., Q_i, ...$, Q_i is denoted as a candidate region , and each candidate region is a minimum circumscribed rectangle of the maximum stable extreme region, and is described by (x, y, w, h) ,where (x, y) are the coordinates of the upper left corner of the rectangle, w indicates the width of the rectangle, h indicates the height of the rectangle. The coincident candidate regions refer to the two candidate regions that are completely coincident, so $(x_i, y_i, w_i, h_i) = (x_j, y_j, w_j, h_j)$. The approximate coincident candidate regions refer to two candidate regions that are not exactly the same, but cover a large common area:

$$Area(Q_i \cap Q_j) \geq k \times Max(Q_i, Q_j) \tag{1}$$

In (1) The Area function calculates the common area of two regions, the Max function calculates the larger area of the two, and k represents the degree of approximation. When k is equal to 1, it means complete coincidence. A nested candidate region refers to a region that is completely located inside another region. The process of the combined suppression strategy is as follows:

a) performing an improved MSER algorithm on the original grayscale image, and generating a set of candidate regions;

b) judging the coincidence degree of the candidate regions in the set. First use the hash algorithm to generate a hash value h_i corresponding to one-to-one correspondence for the quads (x_i, y_i, w_i, h_i) of Q_i , where Q_i represents each of the candidate region sets, then it is judged whether it is coincident by h_i . If it is coincident, it is marked as to be deleted;

c) performing an approximate coincidence determination on an region that is not marked as a duplicate, and marking a candidate region having a larger area in the approximated coincident candidate region as being to be deleted;

d) Approximate nesting judgment is performed on the regions not marked as duplicates, and the nested candidate regions must satisfy the inclusion relationship, and the area and coordinates of the two cannot be too different. If the area of the nested candidate regions is small, the region is marked as to be deleted;

e) Delete all candidate regions marked as to be deleted.

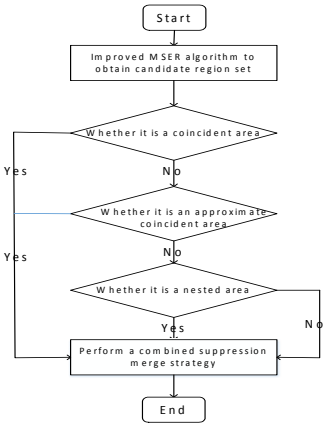

FIGURE IV. FLOWCHART OF COMBINED SUPPRESSION STRATEGY

3) Char-CNN scene text candidate regions classifier

As shown in Figure V, the Char-CNN classifier classifies candidate regions into training and classification categories. The dataset needs to be prepared before training and classification. The original dataset is derived from ICDAR2013. After the improved MSER algorithm, the original dataset will extract many candidate regions, and each candidate region will generate a single image and be normalized into a size of 48 × 48. The improved MSER algorithm is used to perform the maximum stable extremum region detection on the training set of ICDAR2013, and each detected image block is saved as a JPG format image. The training dataset contains a total of 229 natural scene images, and after the training, about 24,000 small images are generated. Then, 1500 character region image blocks are selected from the small images as positive samples, and 8000 typical background region image blocks are selected as negative samples. All images are divided into training datasets, validation datasets, and test datasets. After the dataset is ready, train Char-CNN first.

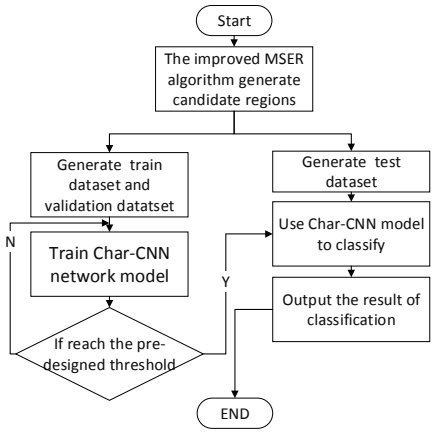

FIGURE V. CANDIDATE REGION CLASSIFICATION PROCESS BASED ON CONVOLUTIONAL NEURAL NETWORK

Char-CNN training process is as follows:

a) Set the framework of the Char-CNN network model. Char-CNN is a 6-layer network. Beside the input layer and the output layer, the intermediate hidden layer contains three convolution layers and one fully connected layer.

b) Initialize the model parameters of Char-CNN, including convolution kernel, bias, activation function, gradient optimization algorithm etc.

c) Train the Char-CNN network model. The training process is batched. The samples are randomly selected from the training set for forward training, and then the loss function value is calculated and the network parameters will be adjusted through the backward propagation;

d) Repeat the process c) until the accuracy of the model reaches the pre-designed threshold or complete all training times.

After completing the Char-CNN training, the candidate regions can be calculated sequentially using the trained Char-CNN model. If the result obtained is 0, the region is labeled as background. If the result obtained by the Char-CNN classifier is 1, it indicates the candidate region is a character region.

B. Character Region Merging Algorithm Using Hierarchical Clustering

This paper adopts a bottom-up condensed hierarchical clustering algorithm. First, each element in the dataset is treated as a single cluster. Then the merge operations are performed on these clusters to form larger and larger clusters, until the expected termination condition is met, or all clusters are merged into one cluster.

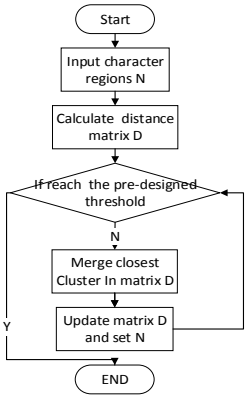

FIGURE VI. HIERARCHICAL CLUSTERING ALGORITHM

The clustering process of character regions in this paper is shown in Figure VI. The process of clustering the character regions is as follows:

a) Each character region is treated as a single cluster, and the distance $d(i, j)$ between the clusters is calculated. The set consisting of $d(i, j)$ is the initialized distance matrix D;

b) Performing a merge operation on the two clusters closest to each other and form a new cluster;

c) Selecting the smallest value among the distances between the new cluster and the original cluster as the similarity between the clusters, updating the inter-cluster distance and the distance matrix D;

d) Repeat steps b) and c) until the inter-cluster distance and intra-cluster distance are optimal, or all clusters are merged into one new cluster.

The final hierarchical clustering algorithm will output clustered K sets, and the centroids of the character regions in each set are considered to be on the same horizontal line.

III. EXPERIMENT RESULTS AND ANALYSIS

ICDAR2013 dataset is chosen to test the algorithm proposed in this paper. Firstly, the performance of the proposed improved MSER algorithm for character candidate regions is tested. The improved MSER algorithm minimizes the detection of the background area while ensuring that the character area is detected.

For the improved MSER algorithm compared with the classic MSER algorithm, the character recall rate increases from 88.9% to 90.2%, and the proportion of detected character

regions increased from 3.25% to 35.19%. The combined suppression strategy can filter out a large number of coincident regions, approximate coincident regions and nested regions.

TABLE I. IMPROVED MSER ALGORITHM VS CLASSIC MSER

Algorithm	Number of character regions detected	Number of non-character regions detected	Character area ratio detected	Character recall rate
Classic MSER	5207	160049	3.25%	88.9%
Improved MSER	5283	15009	35.19%	90.2%

The performance of Char-CNN was evaluated in Table II against SVM classifier based on the LIBSVM software package in OpenCV. The joint feature of HOG and LBP is used as the feature vector of SVM.

TABLE II. ACCURACY OF CLASSIFICATION OF CANDIDATE REGIONS

Classifier	Number of samples	Correct classification number	Accuracy
Char-CNN	21380	20011	93.6%
SVM	21380	19135	89.5%

The accuracy of Char-CNN classification is 93.6%, and the classification accuracy of SVM classifier is 89.5%. The reason is the SVM classifier only uses fixed operators to extract image features, while the Char-CNN classifier uses multiple convolution kernels to iteratively train on multiple convolutional layers and repeatedly optimizes the convolution kernel during continuous training.

The overall performance of the proposed algorithm is evaluated, and compared with other main MSER-based text detection algorithms in recent years in Table III.

TABLE III. MSER-BASED SCENE TEXT DETECTION ALGORITHM

Algorithm	Recall rate	Accuracy	F-Measure
improved MSER algorithm	0.68	0.85	0.76
Ye et al [4]	0.70	0.80	0.74
Tang et al[7]	0.75	0.83	0.79
Neumann et al[9]	0.65	0.74	0.69
Zamberletti et al[10]	0.71	0.80	0.75

Although the algorithms used in the comparison experiments in this paper are based on traditional MSER, the specific implementations are different. The algorithm of Neumann et al. belongs to the earlier scene text detection algorithm, and its recall rate and accuracy are not very high. Ye et al. used the traditional MSER algorithm to detect candidate regions on multi-color channels. Zamberletti et al. used multi-scale MSER algorithm to detect candidate regions. Tang et al. used multiple thresholds for multi-color channels for candidate region detection.

The recall rate of the proposed algorithm for the text region is 0.68, which is higher than that of Neumann's algorithm, but slightly lower than the other three algorithms. The accuracy of the proposed algorithm is 0.85 higher than other algorithms. Because in the candidate region detection phase, this paper only uses the improved MSER algorithm to detect the candidate region on the gray image of the original image, while other algorithms use the MSER algorithm to detect the candidate region in multi-channel or multi-size. However, this paper uses a combined suppression strategy to filter a large number of coincident regions, approximate coincident regions, and nested regions, so that the proportion of character regions in the candidate regions extracted in this paper is much higher than that of other algorithms, which is beneficial to improve the accuracy rate of the candidate region classification phase.

In addition, in the classification stage of candidate regions, this paper uses Char-CNN classifier, which extracts features using multiple convolution kernels, and continuously optimizes convolution kernel parameters in a large number of training processes, and achieves better classification performance, which ultimately leads to a higher overall accuracy of text detection. The algorithm proposed in this paper has also achieved good results on the F-Measure indicator.

IV. CONCLUSION

Based on the existing scene text detection algorithm, this paper combines the improved MSER algorithm and convolutional neural network to detect the scene text. Compared with the same type of algorithms, the text detection algorithm has a competitive comprehensive performance.

ACKNOWLEDGMENT

This work is co-funded by JCYJ20160531194457572, basic research project of Shenzhen science and Technology Innovation Committee, and 2015 R&D support foundation of Shenzhen Virtual University Park: Shenzhen branch of DSSL, project of research and platform construction, and in part by Independent innovation research foundation of HUST.

REFERENCES

[1] K Wang, S Belongie. Word Spotting in the Wild. European Conference on Computer Vision, 2010, 6311:591~604

[2] C Shi, C Wang, B Xiao, el al. Scene Text Recognition Using Part-Based Tree-Structured Character Detection. IEEE Conference on Computer Vision & Pattern Recognition, 2013, 9 (CVPR):2961~2968

[3] Z Zhang, W Shen, C Yao. Symmetry-based text line detection in natural scenes. IEEE Conference on Computer Vision & Pattern Recognition, 2015:2558~2567

[4] Q Ye, D Doermann. Scene Text Detection via Integrated Discrimination of Component Appearance and Consensus. Springer International Publishing, 2013, 8357:47~59

[5] XC Yin, X Yin, K Huang. Robust Text Detection in Natural Scene Images. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36(5):970~83

[6] X Yin, XC Yin, HW Hao. Effective text localization in natural scene images with MSER, geometry-based grouping and AdaBoost. International Conference on Pattern Recognition, 2012:725~728

[7] YB Tang, W Bu, XQ Wu. Natural scene text detection based on multi-level MSER. Journal of Zhejiang University, 2016, 50(6):1134~1140

[8] HI Koo, DH Kim. Scene Text Detection via Connected Component Clustering and Nontext Filtering. IEEE Trans Image Process, 2013, 22(6):2296~2305

[9] L Neumann. Real-time scene text localization and recognition. Computer Vision & Pattern Recognition, 2012, 157(10):3538~3545

[10] A Zamberletti, L Noce, I Gallo. Text Localization Based on Fast Feature Pyramids and Multi-Resolution Maximally Stable Extremal Regions. Asian Conference on Computer Vision, Springer International Publishing, 2014:91~105