Journal on Artificial Intelligence JAI, vol.2, no.1, pp.29-37, 2020

A Method of Text Extremum Region Extraction Based on Joint-Channels

Xueming Qiao[1], Yingxue Xia[1], Weiyi Zhu[2], Dongjie Zhu[3, *], Liang Kong[1], Chunxu Lin[3], Zhenhao Guo[3] and Yiheng Sun[3]

Abstract: Natural scene recognition has important significance and value in the fields of image retrieval, autonomous navigation, human-computer interaction and industrial automation. Firstly, the natural scene image non-text content takes up relatively high proportion; secondly, the natural scene images have a cluttered background and complex lighting conditions, angle, font and color. Therefore, how to extract text extreme regions efficiently from complex and varied natural scene images plays an important role in natural scene image text recognition. In this paper, a Text extremum region Extraction algorithm based on Joint-Channels (TEJC) is proposed. On the one hand, it can solve the problem that the maximum stable extremum region (MSER) algorithm is only suitable for gray images and difficult to process color images. On the other hand, it solves the problem that the MSER algorithm has high complexity and low accuracy when extracting the most stable extreme region. In this paper, the proposed algorithm is tested and evaluated on the ICDAR data set. The experimental results show that the method has superiority.

Keywords: Feature extraction, scene text detection, scene text feature extraction, extreme region.

1 Introduction

The rapid development of wireless communication technology has promoted the rapid spread of the Internet of Things. Widely distributed sensor devices generate large amounts of digital image and video information at all times, and these digital images and video information are mostly natural scenes that describe people's daily lives. Natural scene recognition has important significance and value in the fields of image retrieval [Tsai, Chen, Chen et al. (2011)], autonomous navigation [Rong, Yi, Tian et al. (2016); Zhu, Liao, Yang et al. (2017)], human-computer interaction and industrial automation [Kisacanin, Pavlovic, Huang et al. (2005)] and so on. Compared with traditional text extraction, text extraction in natural scene images faces more severe challenges. First, how to extract the lower text area from the original image is the first challenge because of the non-text content. Secondly,

[1] State Grid Weihai Power Supply Company, Weihai, China.

[2] State Grid Shandong Electric Power Company, Jinan, China.

[3] School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China.

[*] Corresponding Author: Dongjie Zhu. Email: zhudongjie@hit.edu.cn.

Received: 03 February 2020; Accepted: 05 March 2020.

because natural scene images have cluttered backgrounds and complex lighting conditions, angle, fonts and colors, how to make the text detection and ex-traction algorithms immune to these interference factors is another challenge. Therefore, how to extract extreme regions efficiently from complex and varied natural scene images plays an important role in natural scene image text recognition.

So far, there have been many researches in the field of text extreme regions of natural scene images. Shivakumara et al. [Shivakumara, Phan et al. (2010)], Epshtein [Epshtein, Ofek and Wexler (2010)], Lee et al. [Lee, Cho, Jung et al. (2010)] and others use the text area link method to improve the accuracy of the text recognition and the recall rate, but when the selected features are not obviously different between the text and the background, the accuracy of the method will be significantly reduced. In addition, this algorithm is sensitive to image resolution and background interference. Neumann et al. [Neumann and Matas (2010)] proposed the Maximally Stable Extremal Regions (MSER) method, but it can only process gray images and cannot handle color images well; simultaneously, the computational method is complicated when extracting the most stable extreme regions. And the accuracy of the algorithm is not high.

This paper proposes a Text extremum region Extraction algorithm based on Joint-Channels (TEJC). On the one hand, it can solve the problem that the maximum stable extremum region (MSER) algorithm is only suitable for gray images and difficult to process color images. On the other hand, it solves the problem that the MSER algorithm has high complexity and low accuracy when extracting the most stable extreme region. Proved by empirical experiments, the proposed method is superior to the existing methods in feature extraction accuracy and recall rate, and can be well applied to natural scene text detection.

2 Related work

Many excellent algorithms have been proposed previously. Neumann et al. [Neumann and Matas (2010)] proposed the Maximally Stable Extremal Regions (MSER) method. The basic principle of MSER is to binarize a gray image by a threshold, and the threshold is incremented from 0 to 255. The increment of the threshold is similar to the rise of the water surface in the Watershed Algorithm. As the water surface rises, some of the shorter hills will be submerged. If we look down from the sky, the earth is divided into two parts, land and water, which is like a binary image. In all of the binary images obtained, some of the connected regions in the image change little, or even without change, so the regions are called the maximum stable extremum region. However, the MSER algorithm is only applicable to gray images and is difficult to process color images. In addition, the algorithm requires a huge computational cost in the process of finding the maximum stable extreme value region, and it is difficult to achieve fast image text feature extraction. Viswanathan [Viswanathan (2009)] proposed the Features from Accelerated Segment Test (Fast) algorithm. Firstly, a segmentation test is performed on pixels on a fixed radius circle, and a large number of non-feature candidate points can be removed through logic testing; secondly, based on the classification of edge feature detection, the ID3 classifier is used to determine whether the candidate points have edge features according to the 16 features, and the states of each feature are -1, 0, 1; finally, the verification of the corner feature is performed using non-maximum suppression. Because it does not involve scale, gradient,

and other complex operations, Fast detector is very fast. However, its shortcoming is that it has scale invariance and no directionality, and it is not suitable for the characteristics of complex and changeable text direction and scale in natural scenes. Huang et al [Huang, Lin, Yang et al. (2013)] proposed the Stroke Feature Transform (SFT) method. This method is based on Stroke Width Transform, which solves the problem of edge point mismatch in Stroke Width Transform method. Based on this, the factors such as color continuity and local edge point limitation are introduced, which greatly improved the component extraction effect. However, it has a disadvantage that it is only applicable to horizontal texts, and it is not well adapted to the complex and varied character layout in natural scenes.

According to the above analysis, although the existing image feature extraction algorithm has achieved great success in the traditional field, or can solve the problem of character feature extraction in natural scene image recognition to some extent, nowadays, in the face of massive complex and varied natural scene images, how to find a natural scene image text extremum region extraction algorithm with high accuracy and avoiding high complexity calculation is an urgent scientific problem.

3 Extremum region extraction based on joint-channels

This paper proposes an algorithm based on joint-channels for natural scene image text extremum region extraction. Firstly, the algorithm introduces a multi-channel joint method to solve the problem of low character detection rate of single-channel algorithm under the premise of ensuring the operation speed; secondly, on this basis, the multi-channel extremum region extraction algorithm proposed in this paper greatly reduces the computational complexity while ensuring the extraction of the most stable extremum region.

First, we introduce the color spaces.

1). RGB color space

RGB color space, also known as the three primary color light mode, which describes the other colors by three colors: Red, Green, and Blue. We know that the sun is colorless, but there are seven colors that appear after passing through the prism. From here we know that the original colors can be superimposed on each other to get a variety of colors. When the red, green, and blue colors are added in a certain ratio, various colors are obtained, as shown in Fig. 1.

2). HSI color space

The HSI color space is a description method for expressing colors using three basic components of hue, saturation, and intensity. It was proposed by American scientist H.A. Munseu in the 1910s to reflect the different perceptions of color in the human visual system. It can be seen from this that the HSI color space is mainly used to detect and analyze the color characteristics of images. Hue reflects the perception of color in the human visual system, such as yellow, purple and green; Saturation, which indicates the purity of color, the impurity here refers to white light, that is, the more white light mixed in the color, the smaller the saturation value of the color, and the less white light mixed in the color, indicating the purity in color. The greater the degree, the greater the saturation value; I is mainly used to represent the brightness of the image and the energy intensity of the color.

The higher the brightness of the color, the more energy it contains, the greater the intensity I value, and the smaller the intensity I value.

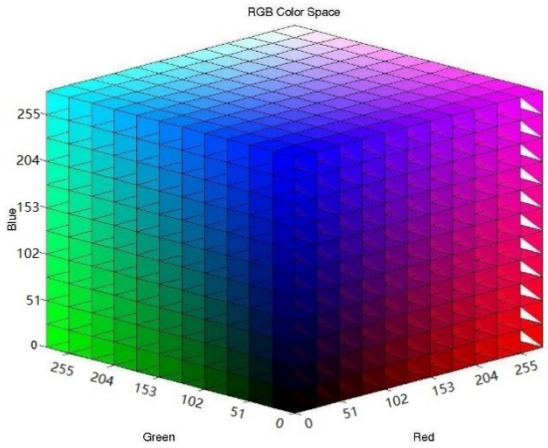

Figure 1: RGB color space

3). Intensity gradient ∇ color space

The intensity gradient ∇ mainly reflects the color relationship between pixels and pixels in the image, showing the gradient trend of color in the image. Each pixel in the image is compared to the neighboring pixels around it, and then the gradient value in which the intensity difference is greatest is given to the pixel. In this way we get the intensity gradient of the image. From the gradient we can see the color gradient of the image.

The color channels involved in this paper are mainly: Saturation (S), indicating the purity of the color; Hue (H), reflecting the perception of color in the human visual system, such as yellow, purple and green; the intensity I is mainly used to express the brightness of the image and the energy intensity of the color, the higher the brightness of the color, the more energy it contains, and the greater the intensity I, and vice versa;

The most stable extreme region (MSER) has high computational complexity and low accuracy. Therefore, we no longer calculate the most stable extremum region, but only extract the extremum region of the image. At the same time, all text areas are included as accurately and comprehensively as possible.

Firstly, we convert the input color image into a gray image under the specified channel. Taking 4 channels (I, H, S, and∇) as an example, each pixel on the image has a corresponding gray value. Secondly, in each gray image, we sequentially take the value of the threshold θ between {0,1,2...255} and obtain the extreme value region corresponding to each threshold. An extreme value region ER having a threshold value is composed of a plurality of extreme value regions ERs having a threshold value of $\theta - 1$ and pixel points having a pixel value of θ .That is r = ($\cup_u \in R_{\theta-1}$)∪($\cup_p \in$ D : C(p) = θ), where$\cup_u \in R_{\theta-1}$ represents a set of extreme regions ERs with a threshold of $\theta - 1$ and $\cup_p \in$ D: C(p) = θ represents pixel points having a pixel value of θ , C(p) represents pixel value .All ERs are built into a component tree. Each component of the tree is an ER. The supervisor-

subordinate relationship of the tree is determined by the inclusive relationship between them. When a new component goes from one gray level to another, the pixel values in this component are the smallest, and are treated as leaf nodes, and the value of the node is set to the corresponding threshold. When we increase the gray threshold, two or more components make up a new component, and the new component becomes the new node of the tree and is the parent of the original node. The algorithm for extremum region extraction based on multi-channel joint is shown as Tab. 1.

Table 1: Extremal region extraction algorithm based on joint-channel

Algorithm: Text extremum region Extraction algorithm based on Joint-Channels (TEJC)

Input: color image I containing text.

Output: extremum region ER under 4 channels.

```
1. Convert the input color image to a gray image It in 4 channels separately;
2. For (each pixel point P in It) {
      Establish each pixel P as a node of the ER tree;
        For (each adjacent point C of pixel P) {
         If (The adjacent point of P is in the image, and it is the node of the extremum
         region tree) {
            Query the roots N(P) and N(C) of the node P and the node C, respectively;
                If(N(P)==N(C)) pass;
                Else if(N(P)>N(C)) {
                   Converging the neighbor node C into the current node P;
                   The root of the neighboring point C points to the root node of the
                   current node P;
                }
         }
         Else {
            Converging the ER component tree represented by the current node P
            into the adjacent node C;
            The root of the current node P points to the root node of the neighboring
            point C;
         }
        }
}
3. Get the root of all ERs in the gray image, and build the ER component tree.
```

Through the above algorithm, we get a large number of extreme regions of color images in four channels, but some of them are obviously not text regions, we can delete them directly.

This can reduce the burden of the subsequent text classification algorithm and reduce the time complexity of the algorithm. The principle of regional clearance is as follows:

1). Clear extreme areas with very large areas. In a real image, if the area occupied by the text is more than half the size of the image, then there is basically no sense of positioning. Therefore, we specify that if the area of the extreme value area is greater than half of the overall area of the image, the ER is cleared.

2). Clear the extreme areas with very small areas. As with the principle of removing the extreme area of the area, if the area of the extreme area is very small and there are very few pixels, then there is no sense of positioning. Therefore, we specify that if the number of pixels in the extreme value area is less than 9, then we clear the ER.

3). Clears extreme regions with very small changes in area in the extreme values contained in each other. Because these extreme areas that differ greatly in area from each other tend to overlap, sometimes they may be noise, and there is no need for repeated classification detection (note that they must be extreme areas that are contained between each other).

In the above, we introduce the extreme region extraction algorithm based on joint-channel joint. As shown in Fig. 2, it transforms the color image into grayscale images on four channels (H channel, I channel, S channel and ∇ channel), and then extract the extreme value in corresponding channel.

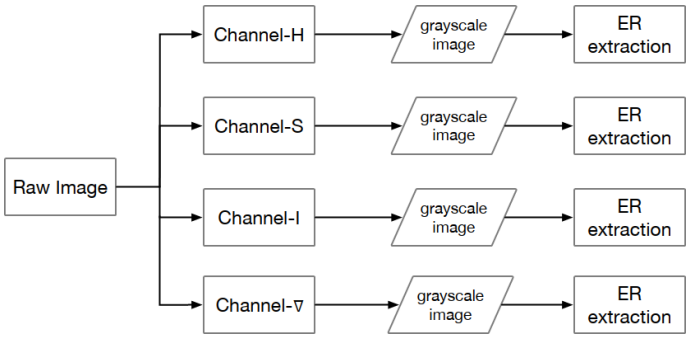

Figure 2: The process of ER extraction algorithm based on joint-channels

This paper uses a simple analogy to demonstrate the ER extraction process.

1) Assume that Fig. 3 is a part of pixels of a gray image.

3	3	3	3	3
3	3	2	3	3
3	2	1	2	3
3	3	2	3	3
3	3	3	3	3

Figure 3: ER Extraction process of initial state

2) Let the threshold increase from 0. The ER obtained first is the ER with threshold 1, which is the red pixel in Fig. 4.

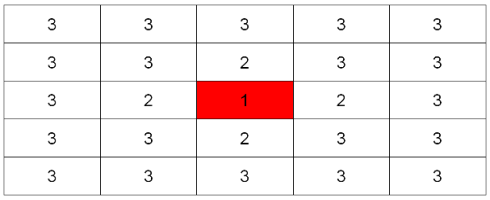

Figure 4: ER extraction process with a threshold of 0

3) The threshold continues to increase, resulting in an ER with a threshold of 2, which is the yellow and red portion of pixels in Fig. 5.

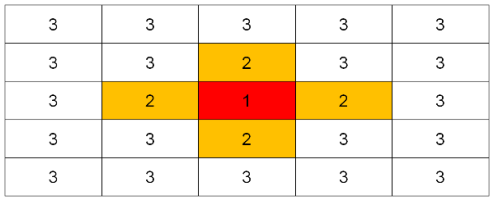

Figure 5: ER extraction process with a threshold of 2

4) The threshold continues to increase, resulting in an ER with a threshold of 3, the green, yellow, and red pixels in Fig. 6.

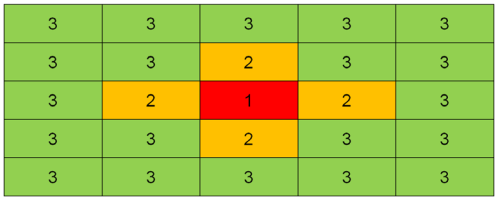

Figure 6: ER extraction process with a threshold of 2

4 Experimental results and analysis

This paper uses the ICDAR 2011 data set [11]: The data set used by ICDAR in the 2011 text-targeting competition contains 255 images including 1189 words and 6,393 letters.

Table 2: Performance of characters detection in the single channel

Channels	Recall (r)	Accuracy (p)
(RGB) R	83.3(%)	7.7(%)
(RGB) G	85.7(%)	10.3(%)
(RGB) B	85.5(%)	8.9(%)
(HSI) H	62.0(%)	2.0(%)
(HSI) S	70.5(%)	4.1(%)
(HSI) I	85.6(%)	10.1(%)

Table 3: Performance of characters detection in multi-channels

Channels	Recall (r)	Accuracy (p)
$I \cup H$	89.9(%)	6.0(%)
$I \cup \mathrm{S}$	90.1(%)	7.2(%)
$I \cup \nabla$	90.8(%)	8.4(%)
$I \cup H \cup S$	92.3(%)	5.5(%)
$I \cup H \cup \nabla$	93.1(%)	5.5(%)
$I \cup R \cup G \cup B$	90.3(%)	9.2(%)
$I \cup H \cup S \cup \nabla$	93.7(%)	5.7(%)
All (7)	94.8(%)	7.1(%)

From the experimental results in Tab. 2 and Tab. 3, we can see that up to 85.7% of the characters in a single channel are detected while the combined detection of characters in all channels can reach 94.8%. However, if the full channel joint detection character is selected, the calculation amount will be large, and the calculation speed will be reduced. Therefore, we can select a reasonable channel combination based on the actual situation after comprehensively considering the character detection effect and the algorithm operation efficiency.

5 Conclusions

This paper proposes an image text localization algorithm based on extremum regions. In the image feature extraction part, we propose an extremal region extraction algorithm based on multi-channel joint to solve the problem that the MSER algorithm cannot process color images well ,and the problem that calculation method has high complexity, and the problem that the MSER algorithm has low accuracy when extracting the most stable extreme regions. However, we also found that replacing the most stable extreme regions with extreme regions will result in more non-text region noise. This can be used to filter the extracted extreme regions through well-trained classifiers. In addition, how to accurately extract features from fuzzy texts, how to explore new image channels, and how to combine them to further improve the accuracy of text positioning are the next studies we are going to do.

Funding Statement: This work is supported by State Grid Shandong Electric Power Company Science and Technology Project Funding under Grant Nos. 520613180002, 62061318C002, the Fundamental Research Funds for the Central Universities (Grant No. HIT. NSRIF.201714), Weihai Science and Technology Development Program (2016DX GJMS15) and Key Research and Development Program in Shandong Provincial (2017GGX90103).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Epshtein, B.; Ofek, E.; Wexler, Y. (2010): Detecting text in natural scenes with stroke width transform. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2963-2970.

Huang, W.; Lin, Z.; Yang, J.; Wang, J. (2013): Text localization in natural images using stroke feature transform and text covariance descriptors. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1241-1248.

Kisacanin, B.; Pavlovic, V.; Huang, T. S. (2005): *Real-time Vision for Human-Computer Interaction*. Springer Science & Business Media.

Lee, S.; Cho, M. S.; Jung, K.; Kim, J. H. (2010): Scene text extraction with edge constraint and text collinearity. *20th International Conference on Pattern Recognition*, pp. 3983-3986.

Neumann, L.; Matas, J. (2010): A method for text localization and recognition in real-world images. *Asian Conference on Computer Vision*, pp. 770-783.

Rong, X.; Yi, C.; Tian, Y. (2016): Recognizing text-based traffic guide panels with cascaded localization network. *European Conference on Computer Vision*, pp. 109-121.

Shivakumara, P.; Phan, T. Q.; Tan, C. L. (2010): A laplacian approach to multi-oriented text detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 412-419.

Tsai, S. S.; Chen, H.; Chen, D.; Schroth, G.; Grzeszczuk, R. et al. (2011): Mobile visual search on printed documents using text and low bit-rate features. *18th IEEE International Conference on Image Processing*, pp. 2601-2604.

Viswanathan, D. G. (2009): Features from accelerated segment test (fast). http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV1011/AV1FeaturefromAcceleratedSegmentTest.pdf.

Zhu, Y.; Liao, M.; Yang, M.; Liu, W. (2017): Cascaded segmentation-detection networks for text-based traffic sign detection. *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 209-219.