2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

A Fusion Strategy for the Single Shot Text Detector

Zheng Yu*, Shujing Lyu*, Yue Lu*, Patrick S. P. Wang†
* Shanghai Key Laboratory of Multidimensional Information Processing
Department of Computer Science and Technology
East China Normal University, Shanghai 200062, China
Email: henuyz@163.com, sjlv@cs.ecnu.edu.cn, ylu@cs.ecnu.edu.cn
† Northeastern University, Boston, MA 02115, United States
Email: patwang@ieee.org

***Abstract*—In this paper, we propose a new fusion strategy for scene text detection. The system is based on a single fully convolution network, which outputs the coordinates of text bounding boxes at multiple scales. We improve the performance of text detection by combining a fusion strategy. This strategy obtains precise text bounding boxes according to the confidence of candidate text boxes. It exhibits promising robustness and discriminative power by fusing text boxes. Experimental results on ICDAR2011 and ICDAR2013 datasets indicate the effectiveness and robustness of the proposed fusion strategy with an F-measure of 87%, which outperforms the base network 2%.**

I. Introduction

Text detection in natural images has become increasingly popular in pattern recognition, computer vision and image understanding. There are numerous applications based on scene text detection, such as license plate location, video caption extraction. Though considerable efforts have been made, scene text detection is still a challenging task due to the unconstrained environment, such as perspective transform, strong light, large-scale occlusion and blurring. Meanwhile, the challenge is also posed by the high variation of text patterns in front, size, color and orientation as well as highly complicated background.

To tackle these challenges, a plenty of approaches have been put forward in recent years. Existing methods are roughly categorized into two groups: (1) hand-designed feature based and (2) deep learning based methods. The hand-designed feature based methods explore various lower-level image properties for text detection. These methods can be further divided into two subgroups: sliding-window and connected component based methods. The sliding-window methods [1], [2], [3] use a multi-scale sub-window to search all possible text in original image. Then a pre-trained classifier is applied to identify whether text is contained within the sub-window. For example, Wang et al. [1] employed a Random Ferns [5] classifier with a histogram of oriented gradients (HOG) feature [6] for text detection. Coates et al. [2] developed an unsupervised feature learning algorithm to generate the feature for text classification. The main difficulties for this group of methods lie in designing a discriminative feature to train a powerful classifier, and managing computational cost by reducing the number of the scanning windows. The connected component based methods are built on bottoms-up strategies, which consist of character candidate extraction, character classification, character refinement, and text line construction in complex images. Stroke Width Transform (SWT) [7] and Maximally Stable Extremal Region (MSER) [8] detectors are two widely used methods in extracting character candidates. However, these methods easily generate false positions of text candidates. This makes it challenging to filter out false detections by a character level classifier.

Recently, the deep learning based methods [9], [10], [11], [12] have been adopted to realize scene text detection. Zhong et al. [9] developed a unified framework to generate word proposal via a fully convolutional neural network (CNN). Gomez-Bigorda et al. [10] proposed a text-specific selective search method that generate a hierarchy of word hypotheses. In addition, Liao et al. [12] proposed a single shot text detector called TextBoxes, which was inspired by an object detector called Singe Shot MultiBox Detector (SSD)[13]. The network of TextBoxes is an end-to-end fully convolutional network, which generates word proposals and directly outputs the coordinates of text bounding boxes at multiple network layers. In order to remove redundancy bounding boxes, a non-maximum suppression algorithm was applied to aggregate all bounding boxes by computing the intersection-over-union (IOU) overlap. However, in scene text detection, the redundant bounding boxes of the letters are remained by non-maximum suppression under the circumstances that the word and letters are detected. The non-maximum suppression algorithm is borrowed from objection detection, but it is not entirely applicable to scene text detection. And the non-maximum suppression only keeps the optimal bounding box when multiple boxes have high IOU overlap. In order to get more accurate detection results, we propose a new fusion strategy for the single shot text detector proposed in [12]. A better text bounding boxes are obtained by this strategy, which fuses text bounding boxes with high confidence. Inspired by recent single deep neural network (i.e. the single shot text detector), the output layers of neural network are used to directly predict text bounding boxes with multi-scale. Then we apply a fusion strategy to aggregate all text bounding boxes. The contributions of this paper is that we provide a more suitable fusion strategy for consolidating text bounding boxes and the proposed fusion strategy is named Text Bounding Boxes Fusion (Text-BBF).

In the rest of this paper, we get details of the proposed method in Section 2. Experimental results and conclusion will

978-1-5386-3787-6/18/$31.00 ©2018 IEEE

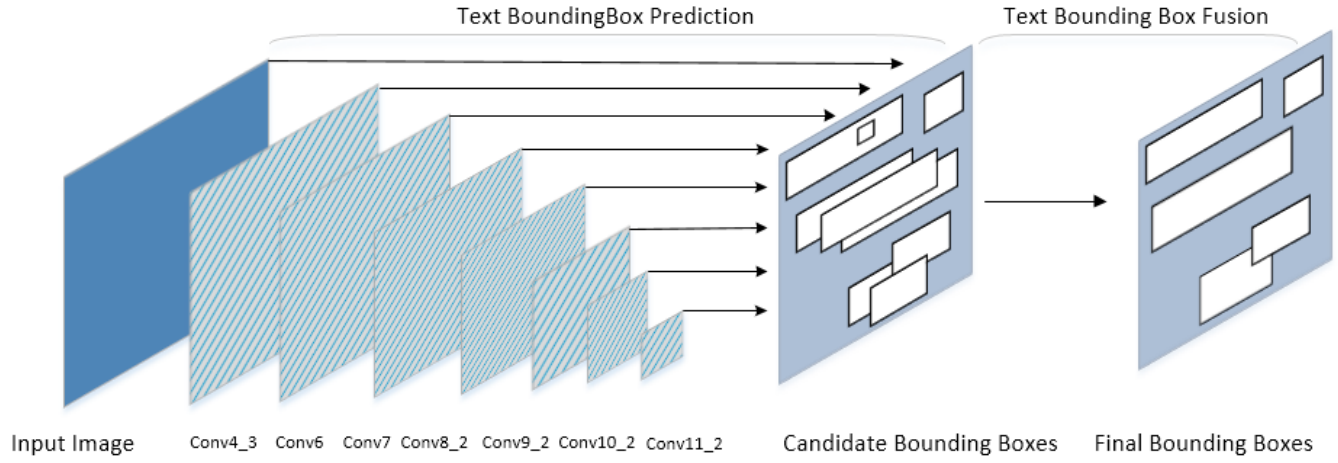


Fig. 1. Structure of the proposed method.

be presented in section 3 and section 4 respectively.

II. Proposed Method

Framework of the scene text detection is shown in Fig.1. Our text detection method consists of two step: 1) Text bounding box prediction. 2) Text bounding box fusion. The text bounding box prediction is based on a single deep neural network called TextBoxes which is proposed in [12]. Multiple output layers of network are applied to predict multiple text bounding boxes. Text bounding box fusion uses an aggregate algorithm to determine the final detection results from all candidate boxes. Note that the proposed text bounding boxes fusion algorithm is only used in the test phrase.

1) Text Bounding Box Prediction: Text bounding box prediction is to predict the bounding boxes of text directly. Based on the capability that TextBoxes[12] predicts bounding boxes as potential text locations and appraises confidence of text categories, TextBoxes is used for predicting text bounding boxes in our research work.

TextBoxes is a fully convolutional network built on the top of a base network of VGG-16 [14] (the last two fully-connected layer are converted into convolution layers). Extra convolution layers are added to the base network and the scale of the convolution layer decreases layer by layer. Because feature maps from different convolution layers have different receptive field sizes, text with various scales can be predicted from multiple feature maps. In addition, multiple aspect ratios are applied to detect text for each feature map. Through combining different scales and aspect ratios from multiple feature maps, a set of candidate text bounding boxes together with corresponding confidence are obtained by this neural network.

2) Text Bounding Boxes Fusion: The output of single neural network is a set of candidate bounding boxes as shown in Fig.1. In the next stage, we try to get precise bounding boxes. In the field of precisely locating object from multiple bounding boxes, non-maximum suppression algorithm is used to address this issue and [12] introduced this algorithm for scene text detection. However, only intersection-over-union (IOU) overlap is used to remove redundancy bounding boxes in non-maximum suppression algorithm. The formula of IOU overlap is as follows:

$$IOU(R_i, R_j) = \frac{Area(R_i \cap R_j)}{Area(R_i \cup R_j)} \tag{1}$$

Where $Area(R_i \cap R_j)$ is the intersection area of rectangles R_i and R_j and $Area(R_i \cup R_j)$ is the union area of rectangles R_i and R_j. $IOU(R_i, R_j)$ denotes the IOU overlap of R_i and R_j.

When a small bounding box of a letter is contained by a big bounding box of a word, the redundant bounding box of the letter is retained by non-maximum suppression in the scene text detection. In addition, the goal of non-maximum suppression is keeping an optimal box when multiple boxes overlap. It dose not use the information of superimposed text boxes.

Therefore, we propose a strategy that fuses multiple text boxes when their IOU overlap and confidence are high. Besides, an extra overlap named inclusion overlap is employed in removing redundant text boxes that are almost contained by other text boxes. Inclusion overlap is defined as follows:

$$I_i(R_i, R_j) = \frac{Area(R_i \cap R_j)}{Area(R_i)} \tag{2}$$

Where $Area(R_i)$ is the area of rectangle R_i and $Area(R_i \cap R_j)$ is the intersection area of rectangles R_i and R_j. $I_i(R_i, R_j)$ denotes the inclusion overlap of R_i relative to R_j.

In order to describe this procedure more clearly, we summarize it as text bounding box fusion algorithm called Text-BBF. The algorithm is presented as Algorithm 1. Considering candidate text boxes set of an image I is $T = t_1, t_2, t_3, ..., t_n$ which is in descending order of confidence and the corresponding confidence set of text boxes is $C = c_1, c_2, c_3, ..., c_n$, we fuse text boxes from T and get a final text boxes set T_{new} that contains the final text boxes of image I.

Algorithm 1 Text Bounding Boxes Fusion (Text-BBF)

Input: text boxes set T and confidence set S
Output: text boxes set T_{new}

while $t_i \in T$ and $t_j \in T(i < j)$ **do**
  if $c_i > \alpha$ and $c_j > \alpha$ **then**
    if $IOU(t_i, t_j) > \beta$ **then**
      $t_i = fusion(t_i, t_j)$
      Set $T \leftarrow T - t_j$
    else if $I_i(t_i, t_j) > \gamma$ **then**
      Set $T \leftarrow T - t_i$
    else if $I_j(t_i, t_j) > \gamma$ **then**
      Set $T \leftarrow T - t_j$
  else
    Set $T \leftarrow T - t_i - t_j$
Set $T_{new} \leftarrow T$
return T_{new}

In Algorithm 1, $IOU(t_i, t_j)$ is intersection-over-union (IOU) overlap of text boxes t_i and t_j, $fusion(t_i, t_j)$ is the merged box of these two boxes. $I_i(t_i, t_j)$ and $I_j(t_i, t_j)$ denote inclusion overlap of text box t_i and text box t_j. There are three thresholds in our algorithm: confidence threshold α, IOU overlap threshold β and inclusion threshold γ. Confidence threshold determines whether two boxes should be fused. There are two situations to integrate text bounding boxes. One is that the IOU overlap of two text boxes is higher than β. we fuse these two text boxes with a merge operation. The merged box is the minimum enclosing bounding box of these two boxes, which replaces the text box with high confidence. Meanwhile, text box with lower confidence is removed. The other case is when the IOU overlap of two text boxes is lower than β and inclusion overlap of a text box is higher that γ. In this case, text box with high inclusion overlap will be removed.

Fig.2 shows the comparison result of text bounding boxes fused by our algorithm and non-maximum suppression algorithm. Fig.2(a) is input image. The candidate text bounding boxes of input image are demonstrated in Fig.2(b). Fig.2(c) exhibits the detection results of non-maximum suppression algorithm and Fig.2(d) shows the detection results with our fusion algorithm (Text-BBF). Note that, in Fig.2(c), the bounding box of letter "W" for the word "WAIT" is retained by non-maximum suppression and the last two lines of text are detected fragmentally although the adjacent text boxes are detected with high confidence. While our method gets a better text detection result that the bounding boxes of text are more complete. It is shown the performance of text detection is improved after using our fusion strategy.

III. Experiments

We evaluate the performance of our proposed method on two scene text detection benchmarks: ICDAR2011 and ICDAR2013 datasets. They are used in the Robust Reading Competition (Challenge 2: Reading Text in Scene Images). The ICDAR2011 dataset contains 229 training images and

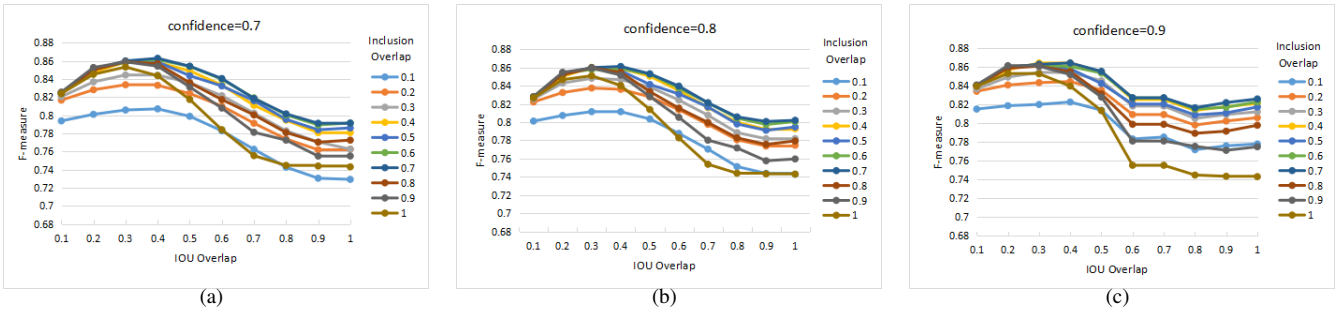
Fig. 2. Example of detection result by non-maximum suppression algorithm and Text-BBF. (a) Input Image. (b) Candidate text boxes with confidence. (c) Text detection result with non-maximum suppression. (d) Text detection result with Text-BBF.

255 test images, while the ICDAR2013 dataset has in total 462 images, including 229 images and 233 images for training and testing, respectively.

In our experiments, a text detector based on TextBoxes is trained for predicting text bounding boxes from input image. The training process is the same as that in [12]. But in the test stage, the method of aggregating candidate text bounding boxes is replaced by our proposed fusion strategy. In order to boost the number of candidate text bounding boxes with high confidence, we rescale input image into multiple scales.

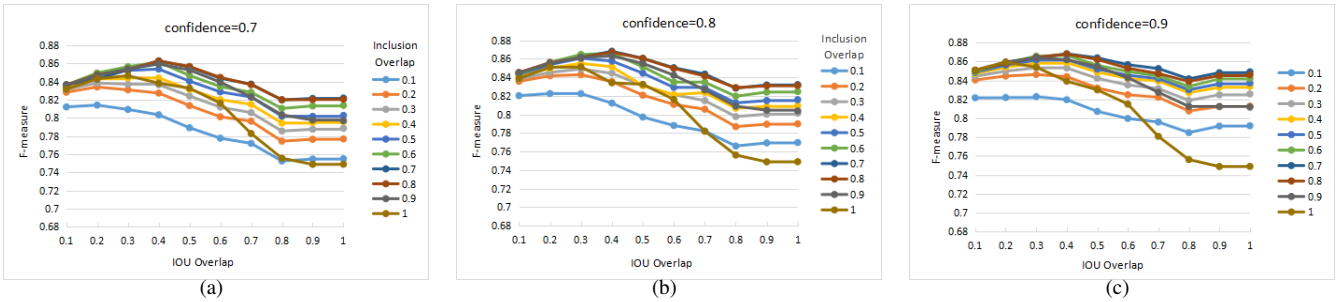
We follow standard evaluation protocol by using the ICDAR 2013 standard [15] on two datasets. To determine optimal thresholds of confidence α, IOU overlap β and inclusion overlap γ for Text-BBF, all the combinations of IOU overlap and inclusion overlap are compared with different value of confidence on ICDAR2011 and ICDAR2013 datasets.

In our algorithm, candidate text boxes with high confidence can be fused. The performances results are compared with three different values of confidence α: 0.7, 0.8 and 0.9. We plot the performance of Text-BBF combined with different kinds of parameter combinations of α, β and γ in Fig.3 and Fig.4. It is obvious that the Text-BBF achieves the best performance when the IOU overlap β is set to 0.4. We compare each combination of β and γ, the optimal combination is β=0.4



(a) (b) (c)

Fig. 3. Comparison of combinations of IOU overlap β and inclusion overlap γ on ICDAR2011 by using different confidence α.



(a) (b) (c)

Fig. 4. Comparison of combinations of IOU overlap β and inclusion overlap γ on ICDAR2013 by using different confidence α.

and γ=0.7 on two datasets. Meanwhile, when α is set to 0.9, our method achieves the best performance and the F-measure can achieve 0.87 on both datasets. In addition, compared to the high performance achieved by Liao el at. [12] on two datasets, our method consistently gives better performance as high as 0.86 with α between 0.2 and 0.5. Through a series of comparison, the confidence α, IOU overlap β and inclusion overlap γ for Text-BBF are set at 0.9, 0.4 and 0.7 respectively in our experiments.

TABLE I
EXPERIMENTAL RESULT ON THE ICDAR2011 DATASET.

Method	Precision	Recall	F-measure
TextBoxes [12]	0.88	0.82	0.85
DeepText [9]	0.85	0.81	0.83
Text Flow [16]	0.86	0.76	0.81
Zhang et al. [3]	0.84	0.76	0.80
MSERs-CNN [11]	0.88	0.71	0.78
Yin et al. [17].	0.86	0.68	0.76
SFT-TCD [18]	0.82	0.75	0.73
Our method	0.90	0.84	0.87

TABLE II
EXPERIMENTAL RESULT ON THE ICDAR2013 DATASET.

Method	Precision	Recall	F-measure
TextBoxes [12]	0.88	0.83	0.85
DeepText [9]	0.87	0.83	0.85
FCN [19]	0.88	0.78	0.83
zhang et al. [3]	0.88	0.74	0.80
Text Flow [16]	0.85	0.76	0.80
iwrr2014 [20]	0.86	0.70	0.77
Our method	0.91	0.84	0.87

The performance of the proposed approach in terms of *Precision*, *Recall* and *F-measure* is shown in Table 1 and Table 2. As we can see, our method shows competitive performance, generally using Text-BBF than non-maximum suppression algorithm. On the ICDAR2011, it outperforms TextBoxes [12] remarkably by improving the F-measure from 0.85 to 0.87. The gains are considerable in both precision and recall, with more than 2% and 2% improvements, respectively. On ICDAR2013 dataset, the Text-BBF improves base network substantially from 0.88 to 0.91 on precision, where the performance of F-measure obtains a 2% improvement. The remarkable significant improvement mainly comes from significantly higher precision by our fusion algorithm (Text-BBF). In addition, we further compare our method against previous methods, it consistently obtains substantial improvements on precision and recall. These results indicate that our fusion strategy is preferred and more principled to solve the problem of scene text detection.

Our detection results on several challenging images are presented in Fig.5. The successful detection samples demonstrate strong robustness against multiple text various and significantly cluttered background. The failure cases have extremely ambiguous text information and have low contrast with its background.

IV. CONCLUSION

We have presented a text bounding boxes fusion strategy for single shot text detector in scene text detection. Firstly, the text detector utilizes multiple feature maps of network to generate candidate text bounding boxes with confidence. Then, the proposed fusion algorithm (Text-BBF) is able to

(a) Successful text detection samples.

(b) Failed text detection samples

Fig. 5. Text detection samples of the proposed method.

handle multiple candidate text boxes robustly. We confirmed that higher accuracy of text detection result is obtained using Text-BBF that fuse bounding boxes with three thresholds: confidence, IOU overlap and inclusion overlap. Experimental results showed that our system achieved the state-of-the-art performance on two standard benchmarks.

REFERENCES

[1] K. Wang, B. Babenko, and S. Belongie, "End-to-end Scene Text Recognition," in*Proceedings of the International Conference on Computer Vision*, pp. 1457-1464, 2011.

[2] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. Wu ,and A. Ng, "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning," in*Proceedings of the International Conference on Document Analysis and Recognition*, pp. 440-445, 2011.

[3] Z. Zhang, W. Shen, C. Yao, and X. Bai, "Symmetry-based Text Line Detection in Natural Scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2558–2567, 2015.

[4] M. Jaderberg , A. Vedaldi, and Z. Andrew "Deep Features for Text Spotting," in *Proceedings of the European Conference on Computer Vision*, vol. 8692, pp. 512-528, 2014.

[5] M. Ozuysal, P. Fua, and V. Lepetit, "Fast Keypoint Recognition in Ten Lines of Code," in*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007,.

[6] N. Dalal, and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 886-893, 2005.

[7] B Epshtein, E Ofek, and Y Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2963–2970, 2011.

[8] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide-baseline Stereo from Maximally Stable Extremal Regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[9] Z. Zhong, L. Jin, S. Zhang ,and F. Feng "DeepText: A Unified Framework for Text Proposal Generation and Text Detection in Natural Images," in *Architecture Science*, vol. 12, pp. 1-18, 2016.

[10] L. Gomez-Bigorda and D. Karatzas "TextProposals: a Text-specific Selective Search Algorithm for Word Spotting in the Wild," in *Pattern Recognition*, vol. 70, pp. 60-74, 2016.

[11] W. Huang, Y. Qiao, and X. Tang, "Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees," in *Proceedings of the European Conference on Computer Vision*, pp. 497-511, 2014.

[12] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast Text Detector with A Single Deep Neural Network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg "Deep Features for Text Spotting," in *Proceedings of the European Conference on Computer Vision*, pp. 21-37, 2016.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *Computing Research Repository*, 2014.

[15] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, Vijay R. Chandrasekhar, and S. Lu, "Icdar 2015 Competition on Robust Reading," in *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 1156–1160, 2015.

[16] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. Tan, "Text flow: A Unified Text Detection System in Natural Scene Images," in *Proceedings of the International Conference on Computer Vision*, pp. 4651–4659, 2015.

[17] X. Yin, K. Huang, and H. Hao, "Robust Text Detection in Natural Scene Images," *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970-983, 2013.

[18] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text Localization in Natural Images using Stroke Feature Transform and text covariance descriptors," in *Proceedings of the International Conference on Computer Vision*, pp. 1241–1248, 2014.

[19] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented Text Detection with Fully Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[20] A. Zamberletti, L. Noce, and I. Gallo, "Text Localization Based on Fast Feature Pyramids and Multi-resolution Maximally Stable Extremal Regions," in *Proceedings of the Asian Conference on Computer Vision*, pp. 91–105, 2014.