Article

Chinese Character Boxes: Single Shot Detector Network for Chinese Character Detection

Junhwan Ryu and Sungho Kim

Department of Electronic Engineering, Yeungnam University, 280 Daehak-ro, Gyeongsan-si, Gyeongsangbuk-do 38541, Korea; ynrjh92@ynu.ac.kr

* Correspondence: sunghokim@ynu.ac.kr; Tel.: +82-53-810-3530

Received: 22 December 2018; Accepted: 15 January 2019; Published: 16 January 2019

Abstract: This paper proposes a deep learning-based Chinese character detection network which is important for character recognition and translation. Detecting the correct character area is an important part of recognition and translation. Previous studies have focused on methods using projection through image pre-processing and recognition methods based on segmentation and methods using hand-crafted features such as analyzing and using features. Unfortunately, the results are vulnerable to noise. Recently, recognition or translation systems based on deep learning were dealt with as a single step from detection to translation but they failed to consider the inaccurate localization problem that arises in detectors. This paper proposes a Chinese character boxes (CCB) network that deals with a method to detect the character area more accurately using the single-shot multibox detector (SSD) as the baseline and called CCB-SSD. The proposed CCB-SSD network has a single prediction layer structure in which unnecessary layers are removed from the feature-pyramid structure. The augmentation method for training is introduced and the problem caused by the use of default boxes is solved by using the proposed non-maximum suppression (NMS). The experimental results revealed a 96.1% detection rate and 0.89 performance against the false positives per character (FPPC) which is the proposed false positive index for the character data-set and caoshu data-set used in this paper. This method showed better performance than the conventional SSD with 69.4% and 6.57 FPPC.

Keywords: deep learning; Chinese character detection; single stage object detector

1. Introduction

Many studies have been conducted to solve optical character recognition (OCR) problems but challenging problems still remain due to various external factors such as the type and boldness of the handwriting, the size of the document, the degree of damage to the document, and the background of the document. In the case of handwritten documents, various noise can occur due to the small size of the text, heavy noise, dense text, and preservation environment. In particular, the preservation environment can cause damage or severe noise. These problems are very difficult to detect and recognize using conventional methodologies. The document covered in this paper is an old document written in caoshu. Caoshu is one of the typefaces of Kanji which corresponds to a curve-oriented cursive and is in the form shown in Figure 1b. Caoshu differs in shape according to the author and has a different form from that of Figure 1a which is the standard Chinese character type today.

Old documents are of high historical value written in a handwritten form and still remain untranslated because of the difficulty and time consuming to translate. It will be a meaningful to discover the history that has not yet been revealed by accelerating the translation of these documents through deep learning which has shown good performance in various fields.

(a) Haesse type example.

(b) Caoshu type example.

Figure 1. Examples of Chinese character typefaces.

This paper proposes a Chinese character boxes (CCB) network for character area detection based on a single-shot multibox detector (SSD) [1] called CCB-SSD to detect each character of an old document. In addition, this paper introduces the detection results based on the number of default boxes and aspect ratios used in SSD and augmentation methods for the cropped data-sets and document type data-sets. Also, a novel non-maximum suppression (NMS) algorithm was proposed to solve the overlap problem caused by the use of several default boxes. A new performance evaluation index, false positives per character (FPPC) was introduced and compared with false positives per image (FPPI) which evaluates the per-image detection performance.

The contributions of paper are as follows.

- CCB-SSD network was proposed for Chinese character detection based on a single-scale structure that can be trained by end-to-end. This can make the network deep enough without additional steps such as deconvolution-based up-sampling and feature map concatenation to overcome the limitations of conventional methods. Therefore, semantic information can be fully exploited and even smaller characters can be detected as a result.
- Using the proposed single-scale structure and NMS, the problem of the existence of unnecessary layers and the relationship between the classifiers was solved and the problem of overlap caused by the use of several default boxes was resolved.
- The FPPC, a new evaluation index that can be a more objective indicator than the FPPI was proposed for the detection of Chinese characters in old documents. In addition, an augmentation method of data-sets with document type and cropped data type was introduced.

The remainder of this paper is organized as follows. Section 2 briefly introduces previous studies related to the detection and recognition of Chinese characters. Section 3 outlines the proposed detector networks called CCB-SSD and training methods. Section 4 introduces various experimental results and finally Section 5 reports the conclusions.

2. Related Works

OCR problems can be divided into deep learning-based methodologies and traditional methodologies before deep learning. The direction of previous research is based on traditional machine learning. Techniques include after image pre-processing the method using projection [2], the recognition method based on segmentation [3–6], method of analyzing the features of the character to use various features [7–15], method to combine words after detecting each letter [16], and recognition and detection method based on the tree structure [17,18]. These conventional methods have a problem in that they are vulnerable to various noises. In particular, they are affected greatly by noise such as spots or blurring.

In recent years, deep learning has been active in many areas of computer vision and deep learning-based methodologies have been explored in the field of character detection, recognition and translation. Tsai et al. [19] used deep convolutional neural networks (CNN) to classify different types of letters. If the network is deep enough for the classification problem, the size of the fully connected layer will not affect the performance. Maidana et al. [20] used four networks, AlexNet [21], VGG [22], LeNet [23], and ZFNet [24], and a single network that was deep enough to assume that the fused network did not have significant performance improvements. Jaderberg et al. [25] first used a network based on R-CNN [26] for detection and a network based on CNN for recognition. Edge Boxes [27] and aggregate channel feature (ACF) detectors [28] were used for the bounding box proposal and the union of the proposed boxes was used as a candidate bounding box in each method. The candidate boxes were then classified by a random forest classifier and finally CNN was used for bounding box regression. The considerable time needed to generate the proposal region by merging the Edge Boxes and the weak ACF detectors and perform all the classification processes using CNN after the regression process is a significant problem. Liao et al. [29] proposed an end-to-end training neural network model for scene text detection. The model uses a fully convolutional structure of 28 layers and is inspired by SSD and uses a single-stage based network. Each layer of the feature pyramid has the same 12 default boxes to predict the same number of outputs, a 72-d vector. The prediction results from each layer are processed through concatenation and NMS, and the processing speed is fast. On the other hand, there is a problem in that it is not well detected when the interval between the characters is large or overexposed. Busta et al. [30] proposed a network that unified text localization and recognition into a single end-to-end network. The detector network was based on YOLO9000 [31] and information about rotation θ was added to the bounding box (x, y, w, h) when the bounding box was predicted. Nevertheless, there is a problem in that it cannot detect a single character or a short snippet of a number and a character.

Networks that existing deep learning-based texts detect by word or sentence. However, an old document written in a caoshu cannot be distinguished from words or sentences because it is created with direction and order from top to bottom and from right to left unlike a typical scene composed of words such as a street sign. Conventional work does not solve this character unit detection problem well.

3. Proposed Method

This section introduces CCB-SSD network which is a network for Chinese character detection and compares the performance according to restrict of random parameters used for data synthesis, and introduces augmentation method of cropped form and document type data. To solve the problem of overlap caused by the use several default boxes and the proposed NMS was used and a new performance evaluation index FPPC was introduced.

3.1. Baseline Method: SSD

Figure 2 presents the structure of conventional SSD. The single-stage detector encapsulates all computations into a single network without a region proposal or feature resampling stage used in a 2-stage detector such as Faster R-CNN [32]. In particular, SSD uses a multi-scale feature map structure to detect objects of various sizes and uses default boxes of various aspect ratios in each feature map to detect objects having various aspect ratios.

Therefore, it is trained to detect small objects in the feature maps of conv4_3 and conv7 which are relatively large and detect large objects in the feature maps of conv9_2 and conv11_2 because the receptive field is large. This paper uses the scale from a minimum of 0.2 to maximum 0.9 and it is necessary to detect objects corresponding to 20% of the input image resolution in conv4_3 which is the largest feature map. On the other hand, conv11_2 is the smallest feature map in which trained to detect an object having a size corresponding to 90% of the input image resolution. By separating the role of each feature map and default boxes, training was made easier.

Figure 2. Network architecture of the conventional SSD.

Three problems typically arise when the SSD structure is applied to the problem of detecting the Chinese character region in the old document treated in this paper. The first problem is that the convolution layer which is predicted in the feature maps of conv7, conv9_2, and conv11_2 in the network structure is an unnecessary layer. The higher layer which is far from the base network is used to detect the objects that occupy most of the image such as cars at close distances but the size of the characters written to the document is not very large. Second, there is a problem of not taking into account the relationship between the results detected from several feature maps. This means that detection can occur multiple times on the same object which can lead to duplicate detection [33]. Finally, using multi-scale feature maps, the convolution layers that detect in the feature map near the base network are lacking in semantic information because the feature map is not deep enough.

3.2. Proposed CCB-SSD Network and NMS

Proposed Network This paper solved the three problems mentioned above by using the proposed single-stage based network and NMS. Figure 3 presents the flow of the whole detection process from the network structure and input to obtain the detection result. Resnet-34 [34] to a $\frac{1}{8}$ scale was used as the base network. The feature map from Resnet-34 passes through 7×3 convolution layers, downsamples to half the size through one 3×3 convolution layer with a stride by 2, and then through $8 \times 3 \times 3$ convolution layers. Therefore, the input scale is downscaled to $\frac{1}{16}$ of the total and the offset and the confidence score for the bounding box are calculated using the 3×3 convolution from the last feature map. Finally, the proposed NMS is used to obtain the final detection results. This is a fully convolutional structure except for one pooling layer of Resnet-34. The orange and blue arrows mean the same feature map.

Training Strategy Training strategies generally follow the content of conventional SSD. Each ground truth box is matched with a default box with best jaccard overlap and default boxes with any ground truth and jaccard overlap greater than 0.5 are then matched. According to conventional SSD, this matching strategy can simplify learning and obtain high scores from the network. The loss function for training is given in Equation (1).

$$L(c, x, l, g) = \frac{1}{N} (L_{conf}(x, c) + L_{loc}(x, l, g)) \quad (1)$$

$$L_{conf} = - \sum_{i \in Pos} x_{ij}^p \log(\hat{c}_i^p) - \sum_{i \in Neg} \log(\hat{c}_i^0) \quad \text{where } \hat{c}_i^p = \text{Softmax} \quad (2)$$

$$L_{loc} = \sum_{i \in Pos} \sum_{m \in \{cx, cy, w, h\}} x_{ij}^k \text{MSE}(\hat{l}_i^m, g_j^m) \quad (3)$$

$$\hat{g}_j^{cx} = \frac{(g_j^{cx} - d_i^{cx})}{d_i^w} \quad \hat{g}_j^{cy} = \frac{(g_j^{cy} - d_i^{cy})}{d_i^h} \quad \hat{g}_j^w = \log\left(\frac{g_j^w}{d_i^w}\right) \quad \hat{g}_j^h = \log\left(\frac{g_j^h}{d_i^h}\right)$$

x is an indicator matrix that indicates whether or not to match. A value of 1 is given if the i -th default box matches the j -th ground truth and 0 otherwise. p is a specific category; N is the number of default boxes matched; l is the predicted box; g is the ground truth box; c is the confidence for category p and d is the default box. The mean squared error (MSE) loss instead of smooth L1 loss was used to obtain the location loss in conventional SSD. Therefore, the MSE function of L_{loc} in Equation (3) is expressed as $(\hat{l}_i^m - \hat{g}_j^m)^2$ and the softmax of Equation (2) is expressed as $\hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)}$. Information about cx , cy , w , and h of the default bounding box and ground truth is encoded and the detector learns information about the difference. Therefore, the coordinates predicted by the detector are offset by the difference. The network initialization method uses He et al. [35] called Kaiming He initialization.

Figure 3. Network architecture of proposed CCB-SSD.

Proposed Non-Maximum Suppression As mentioned above, a conventional SSD uses a multi-scale feature map structure and does not take the relationship between multiple classifiers into account, thereby duplicating the same object. The CCB-SSD network proposed in this paper was a single-scale based detection network but the same problem arises. Each part or stroke of a character also serves as a character. Owing to the nature of these characters, overlapping detection is likely to occur by detecting the regions of actual characters and strokes or other words within them as shown in the right image of Figure 4a. This problem can be solved by using a recognizer, but this paper solves the aforementioned problem in NMS because it is a network study for more accurate detection which is a previous step of recognition. In the case of conventional NMS, confidence is sorted in descending order for a particular class and then repeatedly reduces the overlapping bounding boxes based on the highest confidence. The intersection over union (IoU) is used as a criterion for determining the degree of overlap. To solve the problem shown in Figure 4a, this paper considered the following two areas for the box with the highest confidence and the remaining boxes when calculating the IoU. First, it is important to check if the intersection of the remaining boxes area is greater than 0.6 and to check if intersection of the highest confidence box area is greater than 0.6. Only the box with the highest confidence is left and the remaining boxes are removed. By considering the characteristics of characters unlike ordinary objects,

it is safe to say that the characters do not overlap other characters. Proposed NMS allows to efficiently process both areas without modifying the network or adding additional steps. Figure 4b shows the results of using the proposed NMS. The detection overlapping with the same letter is removed unlike conventional NMS.

Figure 4. Comparison of the results of conventional and proposed NMS for duplicate detection problems.

4. Experimental Results

This section introduces the FPPC and data-set augmentation method, and experiment results according to the parameters of the network.

4.1. Proposed Criterion: False Positives per Character

This section introduces the new evaluation index, false positives per character. The conventional method of calculating the degree of false detection generally uses a FPPI. On the other hand, the data-set and the augmented data-set used in this paper will contain fewer than 10 characters but more often 300 to 400 characters. This is not a problem limited to augmentation but it is also true for actual data-sets. The number of characters written to an image is more accurate than calculating the FPPI because the variation with respect to the number of characters written per image is large. This paper proposes the false positives per character which calculates false positives based on the characters per image as shown in Equation (4). In Equation (4), FA is defined as false alarm ($true\ positives + false\ positives - detected\ GT$) and GT means ground truth. This paper uses 50 characters per image are used, so Equation (4) is multiplied by 50.

$$FPPC = \frac{\# of\ FA}{\# of\ GT} \quad (4)$$

4.2. Kyungpook National University (KNU) Data-Set

The data-set used to train the detector network was the KNU data-set from kyungpook national university. The character data-set contains approximately 200,000 cropped characters and the caoshu data-set consists of 1000 scanned or photographed documents. In addition, there are 100 background images for which there are no letters for augmentation. Table 1 provides an example. The first cropped character data-sets are composed of approximately 200 million data-sets and the second caoshu data-sets are in the form of documents scanned or photographed as cursive texts. The last background data-sets are document-type backgrounds with no text.

Table 1. Three types of data-set examples.

Configuration of data-Set	Cropped Character Data-Set	Caoshu Data-Set	Background Data-Set
# of data-set	200,000	1000	100
example

4.3. Augmentation Method

When a character data-set is augmented, each character is synthesized with the background image. First, the background image is rotated randomly in units of 90 degrees and the characters to be synthesized are fetched. The number of characters in a background image and the size of each character are set to random and the character is resized to a random size with the original ratio maintained. The character is then written while moving randomly along the line. In this case, the boldness of each character is set to a random value and written to the corresponding position. Finally, the method moves along the line and if the position exceeds the image width, it goes to the next line. Therefore, the random parameter for augmentation can be summarized as [background image rotation, number of characters, size of characters, boldness of characters]. Figure 5 shows an example of synthesized images and about 30,000 images were synthesized for training.

When controlling the boldness of the random parameters, if a wide range of random values are used to cover both the boldly written characters and the very faintly written characters, the ground truth is created in empty space as shown in Figure 6. The detector is then trained to detect the empty space based on this information. That is, the accuracy of the detector is adversely affected. When the data-set including this problem and the data-set which restricts the random parameter so that the problem does not occur are trained, the FPPC is reduced significantly from 0.713 to 0.00238 in the same detector. Therefore, these problems may occur when using random parameters.

In the case of a data-set, it is given as a single document type and cannot be used in the above synthesis method, so random cropping is used. A given 1000 documents are cropped randomly and repeatedly to generate approximately 30,000 images similar to the augmented character data-sets. As a result, the image shown in Figure 7 can be obtained.

Figure 5. Example of a character data-set augmented into a background image.

Figure 6. Example of problem caused by an unrestricted random parameter in augmentation.

Figure 7. Example of caoshu data-set augmentation using a random crop.

4.4. Detection Performance Evaluation

Detection performance according to the default box size and ratios The size of the default box specifies the range from the smallest to the largest of the data-sets. The minimum and maximum sizes were set to 10 and 650 respectively. To determine the aspect ratio of the default box, k-mean clustering which was the method used in [31] was applied and the mean IoU according to the number of clusters was compared as shown in Figure 8. Considering the characteristics of the characters written in the document, it is expected that the aspect ratio will be similar. The aspect ratio distribution of clustering using 8 clusters is [1.02, 1.0, 1.03, 1.07, 1.08, 1.6, 1.68, and 1.97] and the actual aspect ratio distribution of the data-set is narrow. On the other hand, a wide range of aspect ratio showed better performance than those with a tight aspect ratio given as a result of clustering and broader aspect ratio. This is because there is no uniformity in that the shape and size of each document vary according to the author. Most aspect ratios of data are within a narrow range but better performance is achieved with fewer default boxes because it does not consider variations in the aspect ratio depending on the author rather than using a wider range of aspect ratios. In addition, a wide range of aspect ratios show good results when actually receiving a new type of data-set as input.

Performance comparison based on the number of layers and whether or not the base network is pre-trained As a base network, Resnet-34 which is most frequently used in detector network is used up to a $\frac{1}{8}$ scale. Because pre-trained models are the trained weights for extracting more general features pre-trained sets are not used to obtain more optimized features to detect Chinese characters. Tables 2 and 3 present their performance. As a result, when using a pre-trained network, there is no direct relationship but the performance improves with increasing number of layers. On the other hand, when using a network that is not pre-trained, the performance is reduced when using more than 16 layers. Although a comparison of the two experimental results did not yield any significant results,

the pre-trained network has relatively fewer FPPC and the pre-trained base network is more stable particularly when the number of layers is small.

Figure 8. Clustering box dimensions on Chinese character datasets.

Table 2. Performance comparison based on the number of layers when not using a pre-trained network.

Fine-Tuning	Character Data-Set		Caoshu Data-Set	
# of Layers	Detection Rate (%)	FPPC (#)	Detection Rate (%)	FPPC (#)
8	97.4	0.27	91.1	10.9
10	99.2	0.46	93.3	8.9
12	98.3	0.27	93.2	8.4
16	99.4	0.27	94.2	8.4
18	99.0	0.27	93.8	8.4
20	99.2	0.45	93.5	10.2

Table 3. Performance comparison based on the number of layers when using a pre-trained network.

Fine-Tuning	Character Data-Set		Caoshu Data-Set	
# of Layers	Detection Rate (%)	FPPC (#)	Detection Rate (%)	FPPC (#)
8	98.2	0.31	92.5	8.9
10	98.9	0.28	93.9	8.5
12	98.3	0.13	93.0	9.1
16	99.1	0.50	94.1	7.4
18	99.4	0.37	93.0	8.7
20	99.4	0.37	94.5	8.0

Performance comparison of random parameter problems Table 4 lists the experimental results for examining the effects of the pixel value threshold on the performance of the augmentation of the data-set mentioned in Section 4.3. When unrestricted random parameters are used, the ground truth is formed in empty space, so the network has a response that needs to be detected. On the other hand, using the restricted random parameter does not cause the criterion of a ground truth in empty space so the same detector is used but the FPPC number is much smaller. Therefore, these problems must be considered in the data augmentation process using random parameters.

Final performance evaluation In Tables 2 and 3, a fine-tuning method was used to compare the performance of the base network in terms of the pre-trained performance. In fine-tuning, the learning rate α is reduced to $\frac{1}{10}$ or $\frac{1}{3}$ of a specific epoch after training. As a result, there is a problem in that the character data-set that has been trained cannot be detected correctly. Therefore, a method of

training together by merging the character data-set and caoshu data-set was used. Table 5 shows the performance of the detector by training the augmentation character data-sets and caoshu data-sets using the restricted random parameter. At this time, the character data-set and caoshu data-set each had approximately 30,000 images and a total of 60,000 images were used for training. The detector performance was evaluated using optimized conditions such as data-set and network structure, NMS, aspect ratio, and size of the default boxes that solved the above-mentioned problems. To confirm the performance based on the 16 layers that showed the best performance in the previous experiment, only 10, 16, and 22 layer deviations were compared. As a result, it showed the best performance with a 96.1% detection rate and 0.89 FPPC when using 16 layers and showed better performance than the conventional SSD with 69.4% and 6.57 FPPC. The detection rate is defined by Equation (5). Figure 9 presents the detection results of conventional SSD and the detection results using the proposed CCB-SSD network. In the case of conventional SSD, the tuning part was set in the same condition as the proposed CCB-SSD network and the entire network structure such as the number of filters and layers was used as the conventional SSD architecture.

$$\text{detection rate (\%)} = \frac{\text{detected GTs}}{\text{number of GTs}} \quad (5)$$

Table 4. Comparison of the effects of the augmentation method on the performance of the result.

Unrestricted Random Parameter		Restricted Random Parameter	
Detection Rate (%)	FPPC (#)	Detection Rate (%)	FPPC (#)
99.4	0.713	99.9	0.00238

Table 5. Performance evaluation using merged data-set and restricted augmentation.

Layers	Detection Rate (%)	FPPC (#)
10	95.20	1.54
16	96.10	0.89
22	89.14	0.74

5. Conclusions

This paper proposed a novel Chinese character boxes-single shot multibox detector (CCB-SSD) network for Chinese character detection. The CCB-SSD network has a single-scale structure which enables the network to be deep enough without any additional processes such as deconvolution-based up-sampling or feature map concatenation for better small character detection. Higher layers in the original SSD were removed because Chinese characters are small and dense. In addition, a new non-maximum suppression (NMS) method was proposed to remove duplicate detections for the same character because the baseline method (SSD) does not consider the relationships between classifiers in the multi-scale feature maps. Finally, a new evaluation index, false positives per character (FPPC) was proposed instead of false positives per image (FPPI) to remove the effect of character numbers per image. Extensive experiments have created an optimal network structure and have shown good performance as a result of testing the network for old documents written in caoshu at an entirely different time. Most characters can be detected well in old documents but there is still a problem of detecting two characters in one character or detecting two characters in one because there are no clear criteria for breaking between characters.

Figure 9. Detection results of conventional SSD and detection results using proposed CCB-SSD. ((a): conventional SSD, (b): proposed CCB-SSD).

Author Contributions: Conceptualization, J.R. and S.K.; methodology, J.R.; software, J.R.; validation, J.R. and S.K.; formal analysis, J.R.; investigation, J.R.; resources, S.K.; data curation, J.R.; writing—original draft preparation, J.R.; writing—review and editing, J.R. and S.K.; visualization, J.R.; supervision, S.K.; project administration, S.K.; funding acquisition, S.K.

Funding: This research received no external funding.

Acknowledgments: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2017M3C1B6071500) and This work was also supported by the 2018 Yeungnam University Research Grants.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin, Germany, 2016; pp. 21–37.
2. Nguyen, K.C.; Nakagawa, M. Text-Line and Character Segmentation for Offline Recognition of Handwritten Japanese Text. *IEICE Techn. Rep.* **2016**, *115*, 53–58.
3. Arica, N.; Yarman-Vural, F.T. Optical character recognition for cursive handwriting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 801–813. [[CrossRef](#)]
4. Roy, A.; Bhowmik, T.K.; Parui, S.K.; Roy, U. A novel approach to skew detection and character segmentation for handwritten Bangla words. In Proceedings of the 2005 Digital Image Computing: Techniques and Applications, (DICTA'05) Queensland, Australia, 6–8 December 2005; p. 30.
5. Tse, J.; Jones, C.; Curtis, D.; Yfantis, E. An OCR-independent character segmentation using shortest-path in grayscale document images. In Proceedings of the Sixth International Conference on Machine Learning and Applications (ICMLA 2007), Cincinnati, OH, USA, 13–15 December 2007; pp. 142–147.
6. Yoon, Y.; Ban, K.D.; Yoon, H.; Kim, J. Blob detection and filtering for character segmentation of license plates. In Proceedings of the 2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP), Banff, AB, Canada, 17–19 September 2012; pp. 349–353.
7. Subramanian, K.; Natarajan, P.; Decerbo, M.; Castanon, D. Character-stroke detection for text-localization and extraction. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, Brazil, 23–26 September 2007; Volume 1, pp. 33–37.
8. Uchida, S.; Shigeyoshi, Y.; Kunishige, Y.; Yaokai, F. A keypoint-based approach toward scenery character detection. In Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, 18–21 September 2011; pp. 819–823.
9. Kunishige, Y.; Yaokai, F.; Uchida, S. Scenery character detection with environmental context. In Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, 18–21 September 2011; pp. 1049–1053.
10. Chen, H.; Tsai, S.S.; Schroth, G.; Chen, D.M.; Grzeszczuk, R.; Girod, B. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In Proceedings of the 2011 18th IEEE International Conference on Image Processing (ICIP), Brussels, Belgium, 11–14 September 2011; pp. 2609–2612.
11. Yi, C.; Yang, X.; Tian, Y. Feature representations for scene text character recognition: A comparative study. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2013; pp. 907–911.
12. Huang, R.; Shivakumara, P.; Uchida, S. Scene character detection by an edge-ray filter. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2013; pp. 462–466.
13. Zhang, J.; Kasturi, R. A novel text detection system based on character and link energies. *IEEE Trans. Image Process.* **2014**, *23*, 4187–4198. [[CrossRef](#)] [[PubMed](#)]
14. Sung, M.C.; Jun, B.; Cho, H.; Kim, D. Scene text detection with robust character candidate extraction method. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 426–430.
15. Zhao, Y.X.; Chou, C.H. Feature Selection Method Based on Neighborhood Relationships: Applications in EEG Signal Identification and Chinese Character Recognition. *Sensors* **2016**, *16*, 871. [[CrossRef](#)] [[PubMed](#)]
16. Wang, K.; Babenko, B.; Belongie, S. End-to-end scene text recognition. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1457–1464.
17. Shi, C.; Wang, C.; Xiao, B.; Zhang, Y.; Gao, S.; Zhang, Z. Scene text recognition using part-based tree-structured character detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2961–2968.

18. Shi, C.; Wang, C.; Xiao, B.; Gao, S.; Hu, J. Scene Text Recognition Using Structure-Guided Character Detection and Linguistic Knowledge. *IEEE Trans. Circuits Syst. Video Tech.* **2014**, *24*, 1235–1250.
19. Tsai, C. *Recognizing Handwritten Japanese Characters Using Deep Convolutional Neural Networks*; Technical Report; Stanford University: Stanford, CA, USA, 2016; pp. 1–7.
20. Maidana, R.G.; dos Santos, J.M.; Granada, R.L.; de Moraes Amory, A.; Barros, R.C. Deep Neural Networks for Handwritten Chinese Character Recognition. In Proceedings of the 2017 Brazilian Conference on Intelligent Systems (BRACIS), Uberlandia, Brazil, 2–5 October 2017; pp. 192–197.
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
23. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *88*, 2278–2324.
24. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin, Germany, 2014; pp. 818–833.
25. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.* **2016**, *116*, 1–20. [[CrossRef](#)]
26. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
27. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin, Germany, 2014; pp. 391–405.
28. Dollár, P.; Appel, R.; Belongie, S.; Perona, P. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1532–1545. [[CrossRef](#)] [[PubMed](#)]
29. Liao, M.; Shi, B.; Bai, X.; Wang, X.; Liu, W. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. In Proceedings of the AAAI, San Francisco, CA, USA, 4–9 February 2017; pp. 4161–4167.
30. Bušta, M.; Neumann, L.; Matas, J. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2231.
31. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
33. Jeong, J.; Park, H.; Kwak, N. Enhancement of SSD by concatenating feature maps for object detection. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 4–7 September 2017.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1026–1034.