ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20241204002168
Original scientific paper

DBNetText Detection Algorithm Based on Edge Detection

Huiqiong FAN*, Changxuan WAN

Abstract: In the booming e-commerce industry, precise text detection in product images is crucial for seamless operations. However, existing text detection algorithms face challenges due to the complex nature of e-commerce images. These images often combine intricate text with complex graphics and diverse product elements, all set against highly variable backgrounds. Artistic fonts, with their unique and often ornate designs, are especially difficult to detect accurately, leading to subpar performance in extracting product-related information. This inefficiency limits the development of intelligent e-commerce applications, which motivates our research. To address these challenges, we propose EIEM-DBNet, an edge-detection-based text detection algorithm. Its key innovation is the integration of the Edge Information Extraction Module (EIEM), which uses operators like Laplace, Sobel, and Canny to extract edge details from low-level feature maps. By emphasizing local edge features, EIEM-DBNet better distinguishes text from the complex background compared to traditional methods that rely on global features. After edge detection, a channel-weighting mechanism incorporates the extracted edge information into the model, enhancing its text detection accuracy. In terms of performance, EIEM-DBNet outperforms traditional DBNet models. In ablation experiments, it shows a 1.1% increase in recall, a 1.3% rise in accuracy, and a 1.2% improvement in F1-score. Compared to other advanced models, EIEM-DBNet achieves the highest recall rate in terms of F1-score, indicating its superior ability to balance precision and recall, thereby providing more accurate text detection in complex e-commerce image scenarios.

Keywords: deep learning; edge detection; multi-scale features; text detection

1 INTRODUCTION

E-commerce images contain a large amount of text information, and in order to successfully extract the information in the images, high-precision text detection and recognition of the images is required [1, 2], after which the extracted text can be semantically analyzed. In fact, e-commerce image text detection belongs to a branch of scene image text detection, and the text in the image has a complex structure and variable background, which makes detection difficult [3].

Recent studies have shown that multi-scale feature fusion can significantly improve the performance of scene text detection. For example, Fan et al. [4] proposed a multi-scale feature fusion module that effectively handles text of varying sizes and orientations by leveraging spatial attention mechanisms to fuse features from multiple levels. Similarly, Huang et al. [5] developed a lightweight model called Efficient Former, which achieves high accuracy while significantly reducing computational costs, making real-time scene text detection more feasible. Moreover, Tang et al. [6] integrated advanced architectures like Transformers into end-to-end scene text detection and recognition systems. These systems utilize multi-scale feature extraction and self-attention mechanisms to enhance detection and recognition accuracy. These advancements highlight the ongoing progress in the field, driven by innovative techniques and architectures.

Early text detection methods mainly used texture-based features [7] and text detection methods based on connected domain component analysis [8]. Texture-based text detection treats text as a special texture and records its strokes for analysis, using the Fourier transform and wavelet transform to separate the text from the background. Text detection based on connected-domain component analysis filters the background pixels based on the position of the text pixels and the proximity of the pixels, and aggregates the text pixels. However, earlier text detection methods could only analyse image information by human selection of features, and such selection of features could not handle complex situations such as artistic fonts and variable backgrounds in e-commerce images.

With the continuous development of deep learning models, academia has gradually developed two types of text detection methods based on deep learning [9]: text detection method based on region regression [10] and text detection method based on pixel segmentation [11]. The text detection method based on region regression will intercept the small images of each target from the input image, and then classify the small images in turn. Text detection tasks usually use the method of preset anchor boxes, that is, for each pixel on the feature map, several text box proposals of different sizes and aspect ratios are preset. The text detection method based on pixel segmentation regards the text detection task as the target segmentation task. Similar to the target segmentation model, the text detection method based on pixel segmentation classifies from the pixel level to determine whether the current pixel belongs to the text target. After discriminating all pixels, the probability map of the text area is obtained, and then the probability map is processed by various post-processing algorithms to obtain the final text detection result map.

Early deep learning text detection methods, such as RCNN [12] and Faster-RCNN [13], initially adapted general object detection frameworks for text detection tasks. While these methods achieved some success, they struggled to effectively handle the unique characteristics of text, such as its elongated rectangular shape and lack of closed contours. For instance, RCNN generates multiple candidate regions and classifies them individually, which is computationally expensive and less effective for text. Faster-RCNN improved efficiency by integrating feature extraction, region proposal, and classification into a single network, but it still lacked the ability to accurately detect text lines due to its reliance on general object detection principles.

To address these limitations, specialized text detection models were developed. CTPN [14], proposed by Tian et al., combined CNN and LSTM networks to

improve horizontal text detection accuracy by introducing a vertical anchor mechanism. Similarly, SegLink [15] detected text segments and combined them into complete text lines, incorporating rotation angle learning to detect multi-oriented text. However, these methods still faced challenges in detecting irregular or artistic fonts, as they primarily focused on horizontal or slightly rotated text.

EAST [16] introduced a significant advancement by enabling the detection of text regions in arbitrary quadrilateral shapes, simplifying post-processing steps and reducing detection time. Despite its efficiency, EAST struggled with curved or highly stylized text due to its reliance on geometric assumptions. PSENet [17] addressed this issue to some extent by using a progressive scale expansion algorithm to distinguish and merge text instances at different kernel scales. However, its pixel-level segmentation approach was computationally intensive and less effective for complex fonts.

LOMO [18] and PAN [19] further improved text detection by incorporating iterative optimization and feature enhancement modules, respectively. LOMO excelled in detecting long text lines and arbitrary shapes, while PAN used feature pyramid enhancement and pixel aggregation to reconstruct text instances. Nevertheless, these methods still faced limitations in detecting artistic fonts, as they primarily relied on geometric and pixel-level features, which are less effective for highly stylized or irregular text.

To solve the mentioned issues above, the CDistNet proposed by Zheng et al. [20] incorporates a multi-domain character distance perception (MDCDP) module, which fuses visual and semantic position embeddings, significantly improving the adaptability to changes in character spacing and direction. On the enhanced dataset containing artistic fonts, the recognition accuracy of CDistNet is 12% higher than that of PSENet, verifying the importance of semantic guidance for the detection of unconventional fonts. Nevertheless, such methods have high demands for computing resources, restricting their application on mobile devices. To address this, Ghosh et al. [21] designed a lightweight CNN-RNN hybrid model. By using MobileNetV2 to compress the parameters to 0.5M, it achieves real time detection while maintaining accuracy, providing new ideas for the embedded application of artistic fonts. It is worth mentioning that the edges detected by existing deep learning methods often exhibit unrefined results and spurious edges. Elharrouss O. [22] employed a cascaded high - resolution network named CHRNet to overcome these challenges.

In summary, while existing methods like CTPN, EAST, PSENet, LOMO, and PAN have made significant progress in text detection, they often fail to effectively detect artistic fonts due to their reliance on geometric assumptions and pixel-level features. A comparative analysis of these models reveals that their performance is highly dependent on the text's regularity and orientation, highlighting the need for more robust approaches that can handle the diverse and complex nature of artistic text.

In terms of application innovation, for instance, regarding the cross-modal retrieval problem in the general domain, a common approach is to adopt pre - trained models and fine-tune them on e-commerce data. Although this approach is straightforward, its performance is less than satisfactory due to the neglect of the uniqueness of e-commerce multi - modal data. The work of Ma et al. [23] has achieved remarkable improvements in addressing this issue. Pan et al. [26] proposed a cross-modal retrieval method based on mixed-scale feature fusion, which optimizes the feature extraction and matching process through advanced neural network architectures. Li et al. [25] further extended this approach to the e-commerce domain, developing a framework that leverages multi-scale feature fusion to significantly enhance retrieval performance.

E-commerce images often exhibit complex text structures and highly variable backgrounds, posing significant challenges for text detection tasks. Traditional text detection models, such as those based on segmentation (e.g., PSENet), have made notable progress in detecting text of arbitrary shapes. However, these models often struggle with the unique characteristics of e-commerce images, such as overlapping text, artistic fonts, and low contrast between text and background. To address these challenges, the DBNet model [26] has emerged as a robust solution. Unlike traditional segmentation-based methods, DBNet introduces a differentiable binarization mechanism that simultaneously predicts a probability map and a threshold map, enabling end-to-end optimization of the binarization process. This approach not only improves detection efficiency but also enhances the accuracy of separating text from complex backgrounds.

Despite its advantages, DBNet has limitations when applied to e-commerce images. Specifically, it relies heavily on high-level semantic features, which may overlook fine-grained edge information crucial for detecting text with irregular shapes or low contrast. This limitation becomes particularly evident in scenarios where text boundaries are blurred or text is embedded in highly textured backgrounds. To bridge this gap, we propose an Edge Information Extraction Module (EIEM) that extracts edge information from low-level feature maps. By integrating edge features with the high-level semantic features used by DBNet, our approach enhances the model's ability to detect text in challenging e-commerce scenarios. The EIEM leverages the spatial details captured in low-level features, which are often lost in deeper layers of the network, to improve the localization and separation of text regions. The main contributions of this paper are summarized as follows:

(1) An edge information extraction module EIEM is proposed. The edge texture information is extracted by edge detection operator, and the edge information is introduced by single channel weighting.

(2) The detection performance of different edge detection operators [27] Sobel, Scharr, Prewitt and Laplace are tested. The detection effects of different text detection models are tested.

2 METHODOLOGY

In this section, we expound upon the DBNet, which has been chosen as the benchmark model for text detection. The overall topological structure of the model is illustrated in Fig. 1. To begin with, we provide a

detailed elaboration of the three key components: feature extraction by the backbone network, feature fusion, and post - processing differential binarization. In next section, we will present a comprehensive and in - depth description of the algorithm of the improved EIEM - DBNET.

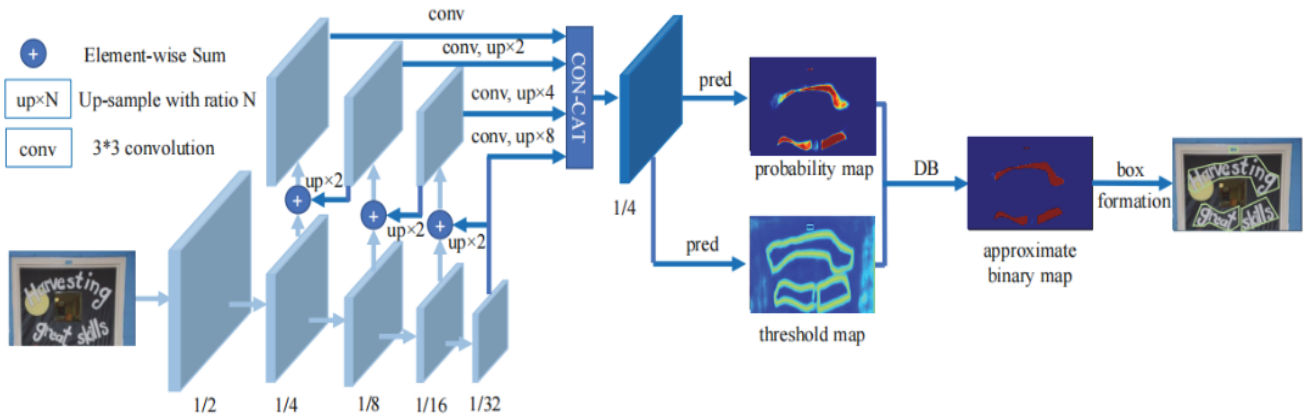

Figure 1 Model structure of DBNet

2.1 Feature Extraction of Backbone Network

The ResNet network is one of the most widely used backbone networks proposed by He [28] et al. in 2015. Its greatest contribution is to solve the 'degradation phenomenon' in neural networks. ResNet has much deeper network layers than other backbone networks. Although in the ideal state, the deeper the number of network layers, the more semantic information is learned, in the actual situation, simply increasing the number of network layers may lead to the phenomenon of 'gradient disappearance' and 'gradient explosion'. Although the above problems are alleviated by increasing the normalization layer, it still cannot change the situation that when the number of network layers increases to a certain extent, the accuracy rate is saturated or even decreases.

In response to this problem, He et al. proposed a residual structure. Assuming that the deep network can simulate any function, when the deep network and the shallow network are identically mapped, the network performance should not be greatly reduced. A reasonable explanation is that when the number of network layers continues to increase, more and more activation functions are introduced, and data mapping has been difficult to return to the origin, that is, with the deepening of the number of network layers, the network has been unable to achieve linear conversion. Therefore, Shortcut Connection is added to the residual connection structure of ResNet. Assuming that the final mapping is $H(x)$, then when the final output of the last shortcut connection is x, the mapping function formula for solving the current network is as follows:

$$F(x) = H(x) - x \tag{1}$$

$F(x)$ is the mapping function of the current network. When the number of network layers is deep and the degradation phenomenon has occurred, it is only necessary to make $F(x) = 0$, then the output of the current shortcut connection is the same as that of the previous shortcut connection, that is, the best output is x, which can ensure that the overall performance of the network does not deteriorate.

2.2 FPN Feature Fusion

In the early target detection network, in order to detect targets of different scales in the image, it is often the case to generate images of different sizes, and generate their own feature maps based on these images, and finally summarize and count the targets detected from the feature maps of different scales. Networks such as Faster-RCNN will predict a single feature map. However, the shallow feature map contains less semantic information, and the location information of the deep feature map is blurred. In order to use the receptive field of the deep feature map and the location information of the shallow feature map at the same time, Lin [29] et al. proposed the FPN structure in 2016. For feature maps of different scales, FPN will first increase the dimension and reduce the dimension through 1×1 convolution, so that the feature map can maintain the same number of channels. The feature map with higher dimensions will undergo multi-layer convolution, so its size will be smaller. For high-dimensional feature maps, bilinear interpolation up-sampling method is generally used to recover their size. After that, feature fusion is completed by splicing or point-by-point addition. DBNet not only uses a top-down approach to obtain a single feature map, but also samples the top-down fused features to the same size and then stitches them again to obtain the final fused features and post-processing.

2.3 Post-Processing Algorithm of DBNet

Text detection method based on pixel segmentation is in the post-processing algorithm and it is often necessary to binarize the predicted feature map. The obtained binary map is connected into a domain by means of aggregation, so that the initial text detection result is obtained. However, the above method has a fatal flaw, that is, the conventional binarization algorithm divides the probability map by a fixed threshold. The formula is as

follows:

$$B_{i,j} = \begin{cases} 1 & \text{if } P_{i,j} \geq t \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

In the above formula, t is the selected probability value. If the overall size of the probability map is $H \times W$, P is any coordinate point in the probability map, and the output in the formula is 1, the pixel on the predicted current coordinate position is a positive sample, that is, text pixel. B is an element in the binarized image matrix, whose value depends on whether P is greater than or equal to a given threshold t. The output of 0 is a negative sample, that is, the background pixel. This binarization method obtains a non-differentiable step signal, which means that the binarization algorithm cannot be optimized during the training process. If you want to identify more text regions, you can only rely on a better feature extraction and feature fusion network.

In order to optimize the post-processing algorithm, Liao et al. proposed a Differentiable Binarization structure. The structure can calculate the binarization probability map through the original probability map and the threshold map to complete the binarization operation. The algorithm can be optimized during the training process, which greatly improves the performance of the text detection algorithm. The formula proposed by Liao et al. is as follows, $\hat{B}_{i,j}$ is the calculated binary probability map, $P_{i,j}$ and $T_{i,j}$ represent the probability map and the threshold map respectively. k is a learning factor, which is set to 50 in the original text. At this time, the signal of the differential binarization algorithm is closest to the step signal of the conventional binarization algorithm.

3 PROPOSED EIEM-DBNet TEXT DETECTION MODEL

Based on DBNet which is elaborated in Section 2 as the benchmark model, we proposed the improved EIEM-DBNet. The improved EIEM-DBNet structure diagram is shown in Fig. 2. Compared with the original DBNet network structure, the feature extraction module is mainly optimized. The Laplace edge detection module is introduced to optimize the original feature extraction module, and the overall performance of the text detection algorithm is improved by extracting the edge features of the text more accurately.

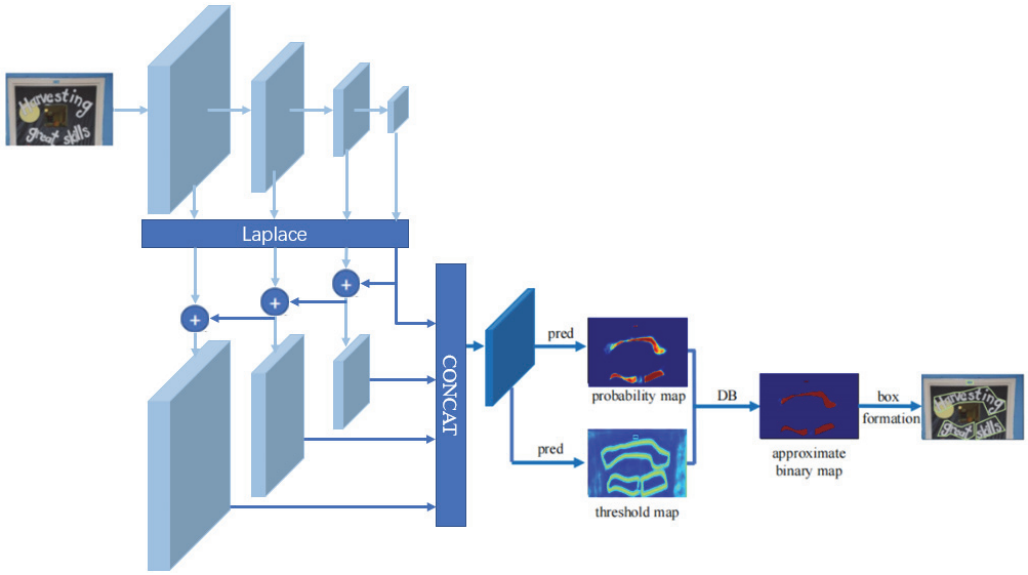

Figure 2 Model structure of EIEM-DBNet

3.1 Feature Extraction of Edge Detection Operator

The text detection method based on pixel segmentation is essentially based on the idea of semantic segmentation to complete the text detection task, which requires the allocation of semantic labels for all pixels. In the DBNet network, the spatial information obtained in the shallow network and the high semantic information in the deep network are used to complete the prediction of the text area through the FPN feature fusion structure. However, in the face of e-commerce images, the method still has some defects. Because the e-commerce image contains more artistic fonts that are quite different from the conventional text, and there are more small texts hidden in the image, the spatial location information extracted from the shallow feature map may not be sufficient, which eventually leads to the problem of text missed detection.

In general, the features in the shallow feature map are more used to detect the spatial information of the text area, but the edge information such as texture features can also be regarded as low-level features. In the field of early text detection, some models use texture-based methods to complete text detection. Text can actually be regarded as a special texture feature. No matter how the art font changes, the color between the font and the font should also be similar. The color between the text and the background is also quite different. Therefore, the edge information of the text is one of the important conditions for distinguishing the text area from the background area, as well as for small-scale text. Therefore, this paper proposes an edge information extraction module to extract edge information from low-level feature maps. The overall structure of the model is shown in Fig. 3.

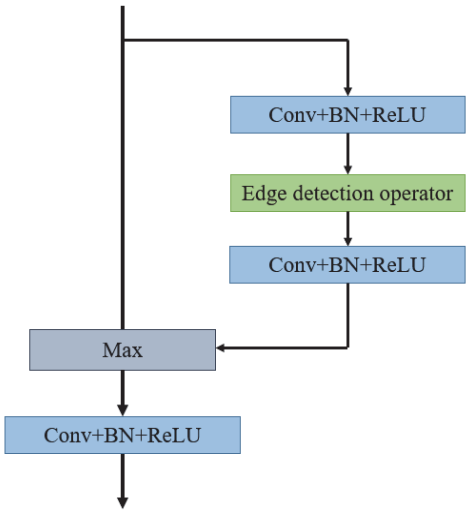

Figure 3 Edge information extraction flow

The shallow feature map has a high resolution, which contains more spatial and edge information of the text area, so the edge detection operator is used to extract the information that is easily overlooked in these original models. The edge information extraction module is directly added to the backbone feature extraction and feature fusion structure, that is, after the feature extraction of the original image is completed, the edge information extraction is introduced before the feature map fusion of different levels. The module adopts a structure similar to the SENet module, and completes the introduction of edge information by single channel weighting. After the feature map is sent to the module, the 3×3 convolution is used to extract the features in the branch, and then the edge information is extracted by using Laplace, Sobel and Canny operators, then, 1×1 convolution dimension reduction is used to adjust the number of feature channels. Then, the feature map of edge information detection is compared with the original feature map pixel by pixel, and the maximum activation value in the same position is obtained, which can detect all possible text areas to the greatest extent. Finally, 3×3 convolution is used to adjust the feature information.

Text regions usually have a large number of discontinuities. The so-called edge refers to the discontinuity of the image in the local area. The edge of the image generally has two characteristics of direction and amplitude, which can be detected by first-order or second-order derivatives. The first-order derivative regards the maximum value as the position of the edge, while the second-order derivative takes the zero point as the corresponding edge position.

Generally speaking, the purpose of edge detection can be completed by detecting the change of gray value. In order to find the direction and amplitude of the edge at (x,y) of the image, it can be achieved by gradient. The specific formula is as follows:

$$\nabla f = \mathrm{grad}(f) = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{3}$$

where ∇f is the gradient, and it also represents a vector, representing the direction of the maximum change rate of f at (x,y). Its size can be expressed as:

$$M(x,y) = \sqrt{g_x^2 + g_x^2} \tag{4}$$

$M(x,y)$ represents the magnitude of the gradient vector at point (x,y), that is, the total rate of change of the pixel value at that point. If you want to calculate all the gradient directions on the feature map, you need to calculate the partial derivative at each position in the map. The specific formula is:

$$\begin{aligned} g_x &= \frac{\partial f(x,y)}{\partial x} = f(x+1,y) - f(x,y) \\ g_y &= \frac{\partial f(x,y)}{\partial y} = f(x,y+1) - f(x,y) \end{aligned} \tag{5}$$

g_x is the gradient value along the x-axis, g_y is the gradient value along the y-axis, and $f(x,y)$ is the pixel value at a given point (x,y) in the image. The calculation of the relevant values of (x,y) can be obtained by filtering (x,y) with one-dimensional template. This filtering module for calculating gradient partial derivative can be called edge detection operator.

The Laplace operator is one of the most common second-order edge detection operators in image processing. Unlike first-order operators such as Sobel and Prewitt, which detect edges by computing the gradient (first derivative) of the image intensity, the Laplace operator calculates the second derivative, making it more sensitive to fine changes in intensity. This operator is rotation-invariant, meaning that changes in the coordinate system do not affect its gradient results, making it particularly useful in applications requiring consistent edge detection regardless of orientation. The Laplace operator is widely used in image enhancement tasks, such as sharpening and noise reduction, due to its ability to highlight regions of rapid intensity change. The template of the operator is:

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{6}$$

The above two templates are 4-neighborhood and 8-neighborhood templates, respectively, representing the gradients in 4 directions and 8 directions around the center pixel of the domain. Through the template, it can be found that when the pixel values around the neighborhood are the same, the gradient can be obtained to be 0, and when the center pixel value is higher than the surrounding pixel value, the gradient is positive, and vice versa. Negative, as long as the convolution result is properly processed, the edge feature detection of the image can be completed.

The Sobel operator is a widely used first-order edge detection operator. It computes the gradient of the image intensity by convolving the image with two 3×3 kernels: one for horizontal changes and one for vertical changes. The Sobel operator is known for its strong noise resistance, making it suitable for applications where the image contains significant noise. However, its ability to detect fine texture features is relatively weak compared to more advanced operators like Scharr. Despite this limitation, the Sobel operator is highly efficient and is often used in real-time applications or scenarios where computational resources are limited.

The Scharr operator shares the same computational efficiency as the Sobel operator and is based on similar principles. However, it improves upon the Sobel operator by amplifying the weight coefficients in its filter kernels. This enhancement allows the Scharr operator to better capture subtle intensity changes, making it more sensitive to fine edges and textures. As a result, the Scharr operator is often preferred in applications requiring high-precision edge detection, such as medical imaging or high-resolution photography. Despite its improved performance, the Scharr operator remains computationally efficient, making it a practical choice for many edge detection tasks.

The Prewitt operator is another first-order edge detection operator that is particularly effective at suppressing noise. Like the Sobel operator, it uses two 3×3 kernels to compute the gradient in the horizontal and vertical directions. However, the Prewitt operator employs a different weighting scheme, which makes it more suitable for images with smooth intensity transitions and significant noise. The Prewitt operator is often used in applications such as industrial inspection and remote sensing, where images may contain noise but require reliable edge detection. While it is less sensitive to fine textures compared to the Scharr operator, its noise suppression capabilities make it a valuable tool in many practical scenarios.

In summary, the four operators-Laplace, Sobel, Scharr, and Prewitt-each exhibit unique characteristics in the context of image edge detection. Considering the specific requirements of this study, our primary focus is on achieving high detection accuracy, while computational speed is of secondary importance. After comparing the various features of different operators, we ultimately selected these four.

4 EXPERIMENTAL SETUP

4.1 Data Set Introduction

ICPR MTWI2018 data set is a network text data set mainly composed of e-commerce images, which contains a variety of fonts and scales of text. The data set was collected and jointly calibrated by South China University of Technology and Alibaba, including a total of 10,000 available images of the label. The difficulty of the data set detection is that the font is complex and changeable, the text pixels cover from one bit to one hundred bits, and there is complex background interference. According to statistics, it is found that some words appear less than 50 times in 10,000 images, which undoubtedly brings great challenges to the detection task. Due to hardware limitations, we randomly selected 3,000 images as the training set, 500 images as the validation set, and 500 images as the test set after filtering out images that did not meet the required specifications. This partitioning ensures a balanced representation of text variations and background complexities across the training, validation, and test sets. To further enhance the model's generalization ability, we applied data augmentation techniques, including random rotation (±15°), horizontal and vertical flipping, and random cropping. These techniques help the model learn robust features from limited data and improve its performance on unseen samples. Fig. 4 lists some images of ICPRMTWI 2018 dataset.

Figure 4 Example images from ICPRMTWI 2018 datasets

4.2 Evaluation Indicators

(1) Accuracy rate

The precision rate can also be regarded as 'precision rate', which is interpreted as 'positive samples detected as positive samples/all samples detected as positive samples', and the formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Among them, TP is the positive sample that is correctly detected, and FP represents the negative sample that is detected as a positive sample.

(2) Recall rate

The recall rate can also be regarded as 'recall rate', which can be interpreted as 'detected positive samples/number of all positive samples'. The formula is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

FN denotes the number of samples that are detected as negative but are actually positive.

(3) F1_score

In fact, precision and recall are a pair of conflicting evaluation indicators. If the model wants to detect more text areas, if its detection performance is not qualitatively improved, it can only rely on detecting more areas. The

increase in the background area misjudged as text will affect the precision rate, and vice versa. In order to comprehensively consider the accuracy rate and recall rate, and avoid being limited to a single detection index, the F1_score is used as a comprehensive index. The formula is as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

4.3 Network Structure and Parameter Setting

In the feature extraction part of the EIEM-DBNet model backbone network, the resnet18 network was selected and pre-trained. The feature input of the FPN feature fusion network was set to 256 dimensions, that is, the feature maps of different scales were unified into 64 dimensions before splicing. The optimizer uses Adam, the initial learning rate is set to 0.001, the batch size is 4, and the number of iterations is unified 80 times. The input feature map is uniformly set at the size of 480 × 480.

5 ABLATION STUDY

Firstly, we analyze the difficulties of e-commerce image detection and the disadvantages of existing models, and know how to optimize the e-commerce image text detection model. Fig. 5 shows the e-commerce image detection map of the traditional DBNet model.

Figure 5 DBNet detection result

(a) Missed detection of small-scale Text. (b) Low-contrast watermark text detection. (c) Background structures misdetected as text: case 1. (d) Background structures misdetected as text: case 2.

It can be seen from Fig. 5 that the normal structure of the text and most of the art fonts have been successfully detected, but there is more small-scale text in the 5a image that is missed, and these texts occupy fewer pixels, so the network can detect from these small texts. There are also fewer text features in the text. In 5b, there are more watermarked texts, which are lighter in color, brighter in background, and lower in contrast between text and background. Therefore, when extracting features, the text features and background features of the watermark text are mixed and difficult to distinguish. In 5c and 5d, there are structures similar to the text on the background, and some structures are mistakenly detected as text. This is also a major difficulty in text detection of e-commerce images. Therefore, it is necessary to use the edge feature extraction module to enhance the text texture features and improve the contrast between the text and the background area.

In order to verify the improvement of the performance of the DBNet text detection model by the edge information extraction module, the ablation experiment was set up. The edge information extraction module was added to the shallow network, the deep network and the full-layer network respectively, and the performance of the respective text detection models was compared.

Among them, the network layer close to the original e-commerce image is regarded as a shallow network, that is, the network whose output of ResNet in the DBNet network is reduced to 1/4 and 1/8 compared with the original image is called a shallow network, and the network whose output is reduced to 1/16 and 1/32 compared with the original image is called a deep network. The full-layer network adds an edge information extraction module in both shallow and deep networks. The model obtained by 80 epoch training is verified on 500 test sets, and the results are shown in Tab. 1.

Table 1 Ablation experiment

Model	Recall / %	Precision / %	F1 score / %
DBNet	55.4	84.4	66.9
DBNet+Laplace (shallow layers)	55.9	86.1	67.7
DBNet+Laplace (deep layers)	55.1	85.4	67.0
DBNet+Laplace (entire layers)	**56.2**	**86.7**	**68.2**

As shown in Tab. 1, after 80 epochs of training, the recall rate of the original DBNet network reaches only 55.4%, while the accuracy rate is 84.4%. This indicates that although the network detects incomplete text areas, it is still effective at distinguishing detected text from the background.

The addition of the Edge Information Extraction Module (EIEM) in the shallow network yields more significant improvements. The recall rate increases by 0.5%, and the accuracy rate improves by 1.7%, demonstrating that the module enhances the network's ability to detect text features more clearly in the shallow layers.

When the edge detection module is added to the deep network, the recall rate decreases by 0.3%, but the accuracy rate improves by 1.0%. This can be explained by the fact that the deep network primarily learns high-level semantic information from the e-commerce images. After applying the edge detection operator to the high-level features, the originally recorded information is disrupted, leading to fewer detectable text areas and a

decrease in recall. However, the reduction in misdetected areas results in an improvement in accuracy, which measures the proportion of correctly detected text areas relative to all detected areas.

Adding the edge detection module to both shallow and deep layers yields the best performance of the detection model. The recall rate increases by 0.8%, accuracy rises by 1.3%, and the harmonic average improves by 1.3%. This is due to the feature fusion module, which combines the shallow network's texture features with the deep network's semantic information. The simultaneous addition of the edge detection module enhances both feature types during training. The changes in average loss and harmonic mean are shown in Figs. 6 and 7. From Fig. 6, it can be seen that the average loss value of the model decreases progressively with the number of epochs. After 40 epochs, the loss stabilizes around 0.6, with a slower decline rate thereafter. By 70 epochs, the loss value stabilizes. The F1-score curve in Fig. 7 shows that adding the edge detection module to the deep feature extraction network impacts the high-level semantic information. Initially, at epoch 1, the F1-score is only 22.2, the lowest stability compared to other models. However, when the edge detection module is added to both the shallow and deep layers, the harmonic average starts higher and remains more stable than other models. In subsequent iterations, the model reaches its maximum F1-score at around 70 epochs and tends to stabilize.

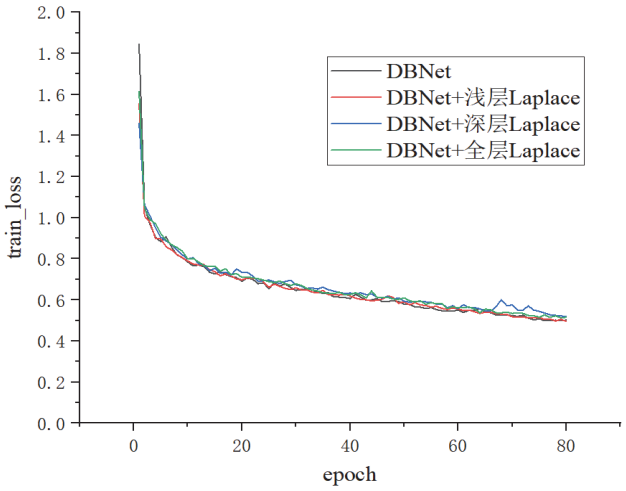

Figure 6 train_loss change curve of DBNet

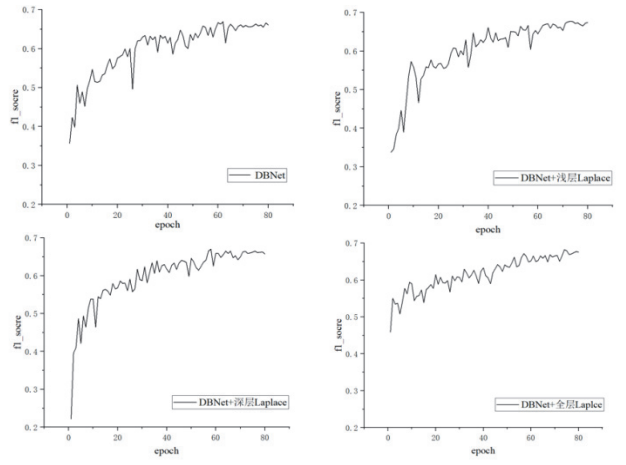

Figure 7 F1_score change curve of DBNet

Fig. 8 shows the comparison of the text detection results of the model after receiving the optimization of the edge detection operator. The first column is the original DBNet model detection diagram. It can be seen that the trademark is mistakenly detected as text in 8a, while it is correctly detected as the background area in 8b. 8d is more complete than 8c small-scale text detection. The long text below 8f can be detected as a complete text line, and the detection result is more accurate.

Figure 8 Comparison of EIEM-DBNet model test results

(a) Trademark misidentified as text. (b) Trademark correctly identified as background. (c) Incomplete detection of small-scale text. (d) Enhanced detection of small-scale text. (e) Detection result of long text. (f) Complete detection of long text line

Multiple edge detection operators have been developed in the field of image processing. In order to verify the performance of different edge detection operators, other edge detection operators are selected for comparative experiments. The edge detection module in the experiment is still designed in both shallow and deep networks. In this section, Sobel operator, Scharr operator, Prewitt operator and Laplace operator in the module are selected for comparison. The results of ablation comparison experiments are shown in Tab. 2. The templates of the above operators are:

$$Sobel_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} Sobel_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (10)$$

$$Scharr_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix} Scharr_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix} \quad (11)$$

$$Prewitt_x = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} Prewitt_y = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad (12)$$

Table 2 Comparison of evaluation results of different edge detection operator

Model	Recall / %	Precision / %	F1_score / %
DBNet	55.4	84.4	66.9
DBNet+Sobel	56.3	85.0	67.7
DBNet+Scharr	56.2	86.1	68.0
DBNet+Prewitt	55.6	86.7	67.7
DBNet+Laplace	**56.2**	**86.7**	**68.2**

After using Sobel operator and Scharr operator, the recall rate is obviously improved, but compared with Scharr operator, Sobel operator has no advantage in accuracy. The mechanism of the two is the same, but the Sobel operator is mainly anti-noise ability, and is not sensitive to fine texture features, so its accuracy rate is only increased by 0.6%. The Scharr operator amplifies the overall weight coefficient, and the edge feature extraction performance of the image has been greatly improved. The accuracy rate has reached 86.1%, and the performance is better than the Sobel operator. The Prewitt operator can reach the optimal level in terms of accuracy, but the computational cost of the operator is high. Therefore, in comparison, the Laplace operator has achieved a balance between recall and accuracy and is the optimal choice.

6 COMPARATIVE EXPERIMENTS

In order to fully illustrate the superiority of EIEM-DBNet, under the premise of ensuring the same number of training sets and test sets as other text detection models, 9000 images are selected for training and 1000 images are tested. This section compares the model with other text detection models. Tab. 3 shows the results of the comparative test. The principles of CTPN, EAST, PSENet and other networks have been stated in this chapter. Faster-RCNN is an early network that achieves high-precision object detection through second-order networks and RPN. It is also a target detection network based on region regression, but the target of this detection is text. The CTPN model can only detect text in the horizontal direction, and the detection effect is not good when facing text with vertical or rotating angles, and the F1_score is only 45.5%. The detection performance of Faster-RCNN in the face of small text is not good, and the EAST model is not good for long text detection. However, due to its multiple extraction and fusion of features, the overall performance is still better than the Faster-RCNN model. On the whole, the performance of the text detection model based on pixel segmentation is better than that of the text detection model based on region regression. In order to highlight the effect of the edge detection operator, this chapter also uses TextRay [30] for an ablation study. This model can perform top-down contour-based geometric modeling and geometric parameter learning within a single-shot anchor-free framework and can detect text of any scale, shape, and orientation. From PSENet, DBNet, TextRay and the model in this chapter, it can be seen that EIEM-DBNet achieves a balance between recall rate and precision rate after feature extraction by edge detection operator, and F1_score achieves the best effect.

Table 3 Comparison of different text detection models

Method	Recall / %	Precision / %	F1_score / %
CTPN	38.5	55.7	45.5
Faster RCNN	51.4	60.3	55.5
EAST	55.6	65.4	60.1
PSENet	60.5	75.6	67.2
DBNet	59.3	88.1	70.8
TextRay	60.0	86.94	71.0
EIEM-DBNet	**61.2**	**87.0**	**71.5**

7 CONCLUSIONS

This paper proposes a text detection model EIEM-DBNet combined with edge detection operators. Firstly, we compare the current commonly used text detection algorithms and analyze the characteristics of e-commerce image text. Then, we analyze and select the basic model DBNet for this chapter. The DBNet model is examined, and its feature extraction module, feature fusion module, and post-processing algorithm module are summarized. An edge information extraction module (EIEM) is proposed, and several different edge detection operators are introduced. Relevant experiments are conducted to analyze the impact of the edge information extraction module's position and the influence of different edge detection operators on text detection. The improved EIEM-DBNet text detection model is compared with other commonly used models, demonstrating significant improvements in recall rate and accuracy rate, and overall model stability.

Our preliminary research has focused on the theoretical analysis of practical applications in e-commerce, discussing this at a theoretical level. In the future, we may apply our findings to real e-commerce platforms such as Taobao, Tmall, JD.com, and Pinduoduo. However, transitioning from theoretical research to practical application requires substantial additional work, including algorithm optimization, construction of diverse datasets, integration with backend services, frontend display integration, and considerations regarding privacy and security.

Moreover, future research will extend to broader application scenarios, including but not limited to text image detection in medical images. These new application contexts demand enhanced noise resistance and computational efficiency from our models. Noise reduction techniques in medical imaging play a crucial role in improving image quality, assisting in diagnosis, and treatment. Therefore, our next research focus will be on enhancing the models' noise resistance and computational efficiency. Additionally, we must address privacy and security concerns to ensure that our technology complies with relevant regulations and standards.

8 REFERENCES

[1] Lari, H. A., Vaishnava, K., & Manu, K. S. (2022). Artificial intelligence in E-commerce: Applications, Implications

and Challenges. *Asian Journal of Management*, *13*(3), 235-244. http://doi.org/10.52711/2321-5763.2022.00041

[2] Jing, J. F., Liu, S. J., Wang, G., Zhang, W. C., & Sun, C. M. (2022). Recent advances on image edge detection: A comprehensive review. *Neurocomputing*, *503*, 259-271. https://doi.org/10.1016/j.neucom.2022.06.083

[3] Xu, P., Huang, S., & Wang, H. (2019). A multi-oriented Chinese keyword spotter guided by text line detection. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, *2019*. https://doi.org/10.1109/ICDAR.2019.00112

[4] Fan, J., Bocus, M. J., & Hosking, B. (2021). Multi-scale feature fusion: Learning better semantic segmentation for road pothole detection. *2021 IEEE international conference on autonomous systems (ICAS)*, *2021*, 1-5. https://doi.org/10.1109/ICAS49788.2021.9551165

[5] Huang, M., Liu, Y., & Peng, Z. (2022). Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, *2022*, 4593-4603. https://doi.org/10.1109/CVPR52688.2022.00455

[6] Tang, J., Zhang, W., Liu, H. et al. (2022). Few could be better than all: Feature sampling and grouping for scene text detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *2022*, 4563-4572. https://doi.org/10.1109/CVPR52688.2022.00452

[7] Naosekpam, V. & Sahu, N. (2022). Text detection, recognition, and script identification in natural scene images: A Review. *International Journal of Multimedia Information Retrieval*, *11*(3), 291-314. https://doi.org/10.1007/s13735-022-00243-8

[8] Liu, S., Xian, Y., & Li, H. (2017). Text detection in natural scene images using morphological component analysis and Laplacian dictionary. IEEE/CAA *Journal of Automatica Sinica*, *7*(1), 214-222. https://doi.org/10.1109/JAS.2017.7510427

[9] Naiemi, F., Ghods, V., & Khalesi, H. (2022). Scene text detection and recognition: a survey. *Multimedia Tools and Applications*, *81*(14), 20255-20290. https://doi.org/10.1007/s11042-022-12693-7

[10] He, W., Zhang, X. Y., & Yin, F. (2018). Multi-oriented and multi-lingual scene text detection with direct regression. *IEEE Transactions on Image Processing*, *27*(11), 5406-5419. https://doi.org/10.1109/TIP.2018.2855399

[11] Xu, Y., Wang, Y., & Zhou, W. (2019). Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing*, *28*(11), 5566-5579. https://doi.org/10.1109/TIP.2019.2900589

[12] Wu, Y., Liu, W., & Wan, S. (2021). Multiple attention encoded cascade R-CNN for scene text detection. *Journal of Visual Communication and Image Representation*, *80*, 103261. https://doi.org/10.1016/j.jvcir.2021.103261

[13] Zhong, Z., Sun, L., & Huo, Q. (2019). An anchor-free region proposal network for Faster R-CNN-based text detection approaches. *International Journal on Document Analysis and Recognition (IJDAR)*, *22*, 315-327. https://doi.org/10.1007/s10032-019-00335-y

[14] Tian, Z., Huang, W., & He, T. (2016). Detecting text in natural image with connectionist text proposal network. Computer Vision-ECCV 2016: 14th European Conference, Proceedings, Part VIII 14, 56-72. https://doi.org/10.1007/978-3-319-46484-8_4

[15] Shi, B., Bai, X., & Belongie, S. (2017). Detecting oriented text in natural images by linking segments. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2550-2558. https://doi.org/10.1109/CVPR.2017.371

[16] Zhou, X., Yao, C., & Wen, H. (2017). EAST: An efficient and accurate scene text detector. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 5551-5560. https://doi.org/10.1109/CVPR.2017.283

[17] Wang, W., Xie, E., & Li, X. (2019). Shape robust text detection with progressive scale expansion network. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9336-9345. https://doi.org/10.1109/CVPR.2019.00956

[18] Zhang, C., Liang, B., & Huang, Z. (2019). Look more than once: An accurate detector for text of arbitrary shapes. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10552-10561. https://doi.org/10.1109/CVPR.2019.01080

[19] Wang, W., Xie, E., & Song, X. (2019). Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. *Proceedings of the IEEE/CVF international conference on computer vision*, 8440-8449. https://doi.org/10.1109/ICCV.2019.00853

[20] Zheng, T., Chen, Z., & Fang, S. (2024). Cdistnet: Perceiving multi-domain character distance for robust text recognition. *International Journal of Computer Vision*, *132*(2), 300-318. https://doi.org/10.1007/s11263-023-01880

[21] Ghosh, J., Talukdar, A. K., & Sarma, K. K. (2024). A light-weight natural scene text detection and recognition system. *Multimedia Tools and Applications*, *83*(3), 6651-6683. https://doi.org/10.1007/s11042-023-15696-0

[22] Elharrouss, O., Hmamouche, Y., & Idrissi, A. K. (2023). Refined edge detection with cascaded and high-resolution convolutional network. *Pattern Recognition*, *138*, 109361. https://doi.org/10.1016/j.patcog.2023.109361

[23] Ma, H., Zhao, H., & Lin, Z. (2022). EI-CLIP: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18051-18061. https://doi.org/10.1109/CVPR52688.2022.01752

[24] Pan, X., Xie, X., & Yang, J. (2025). Mixed-scale cross-modal fusion network for referring image segmentation. *Neurocomputing*, *614*, 128793. https://doi.org/10.1016/j.neucom.2024.128793

[25] Li, T., Kong, L., & Yang, X. (2024). Bridging modalities: A survey of cross-modal image-text retrieval. *Chinese Journal of Information Fusion*, *1*(1), 79-92. https://doi.org/10.62762/CJIF.2024.361895

[26] Liao, M., Zou, Z., & Wan, Z. (2022). Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 919-931. https://doi.org/10.1109/TPAMI.2021.3155612

[27] Al-Taie, R. R. K., Saleh, B. J., & Salman, L. A. (2021). Image edge-segmentation techniques: A review. *International Journal of Scientific Research in Science, Engineering and Technology*, *8*(5), 252-257. https://doi.org/10.32628/IJSRSET218528

[28] He, K., Zhang, X., & Ren, S. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778. https://doi.org/10.1109/CVPR.2016.90

[29] Lin, T. Y., Dollár, P., & Girshick, R. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[30] Wang, F., Chen, Y., & Wu, F. (2020). Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection. *Proceedings of the 28th ACM international conference on multimedia*, 111-120. https://doi.org/10.1145/3394171.3413819

Contact information:

Huiqiong FAN
(Corresponding author)
1) School of Information Management,
Jiangxi University of Finance and Economics,
Nanchang, 330032, Jiangxi Province, China
2) Jiangxi Key Lab of Data and Knowledge Engineering,
Jiangxi University of Finance and Economics,
Nanchang, 330013, Jiangxi Province, China
E-mail: fhq8109@163.com

Changxuan WAN
1) School of Information Management,
Jiangxi University of Finance and Economics,
Nanchang, 330032, Jiangxi Province, China
2) Jiangxi Key Lab of Data and Knowledge Engineering,
Jiangxi University of Finance and Economics,
Nanchang, 330013, Jiangxi Province, China
E-mail: wanchangxuan@263.net