Journal of Database Management
Volume 34 • Issue 1

Text Detection Model for Historical Documents Using CNN and MSER

Rankang Li, Southwest University, China
Shanxiong Chen, Southwest University, China*
Fujia Zhao, Southwest University, China
Xiaogang Qiu, Southwest University, China

ABSTRACT

This article introduces a text detection model for historical documents images. The handwritten characters in historical documents are always difficult to detect because they contain fuzzy or missing ink, or weathering features and stains; these features will seriously affect the detection accuracy. In order to reduce the influence mentioned above, an effective ATD model is proposed to detect the textbox of characters in historical documents image, and ATD model includes a CNN-based text-box generation network and an NMS-based MSER text-box generation model. As a post-processing method, a text merging algorithm is proposed to achieve higher detection accuracy. The test results on historical document datasets such as Yi, English, Latin, and Italian datasets show that the method in this paper has good accuracy, and it has taken a solid step for the detection of historical documents.

KEYWORDS

Convolutional Neural Network, Maximum Stable Extremum Region, Text-box mering, Text Detection of Historical Documents

INTRODUCTION

Handwritten text detection is an important research in computer vision and pattern recognition, and it refers to the task of determining the exact position of all texts or characters from input image and marking it with colored text-box. The difference of writing, outline and shape of handwritten texts made it very difficult to be detected accurately. Therefore, the detection of handwritten text ushered in a difficult challenge. Handwritten text detection has a wide range of applications such as documents recognition, historical documents translation, robot vision, etc. So, it is very important to continuously conduct in-depth research on detection methods in order to improve detection performance.

In the detection task of handwritten characters, images are always slanted, which has defects, ambiguities and excessive background noise, and historical document images have additional problems

DOI: 10.4018/JDM.322086

*Corresponding Author

such as stains and breakage. As a special issue of handwritten text detection, text detection of historical document is performed on historical documents. With the fewer and fewer experts and scholars pay attention to the translation and understanding of historical documents, the importance of an automatic recognition system for historical documents is self-evident. The advantages of historical documents automatic recognition system are as follows: First, all historical documents exist in the form of digital images, avoiding the gradual disappearance such as fading paper or oracle. Second, it can quickly and automatically detect and recognize the input image, which is more efficient and accurate than manual. Third, an efficient detection and recognition system can facilitate the relevant learning of historical documents for researchers. Historical documents detection has an important application in the historical documents recognition system. Because an effective historical document recognition system needs to accurately detect the text-box before it can be recognized.

In recent years, especially with the popularity of deep learning technology, the field of text detection has attracted extensive attention of computer researchers. However, most researchers only focus on scene text detection, document text detection, handwritten text detection and other hot areas. As a special text detection task, the text recognition task of historical document images are difficult to detect because of its complex background, incomplete and fuzzy text, and its initial application value is small, so it has not been paid attention to by the academic community.

In the past 20 years, researchers have proposed many algorithms for text detection in handwritten characters. Especially in the past 10 years, the following literatures are dedicated to the detection of handwritten text (Shin H C, Roth H R, & Gao M, 2017), text detection tasks are defined as a two steps task: Candidate text area extraction and text/non-text area. These algorithms can generally be divided into two categories based on traditional algorithms and algorithms based on deep learning (Chen Shanxiong, Han Xu, & Mo Bofeng, 2017; Chen Shanxiong, Wang Xiaolong, & Wang Minggui, 2019). Because text detection of historical documents is special issue of handwritten text detection, so the methods used in handwritten text detection is suitable for historical documents text detection theoretically. However, additional influence of historical documents image makes it more difficult to detect accurately than handwritten text detection. Therefore, it is very important to improve the existing methods to achieve higher accuracy.

In order to alleviate declines caused by background noise on detection accuracy while maintaining detection efficiency, early research on historical document images detection mostly use traditional algorithms, or texture connected area algorithms which regarded text as a certain specific texture feature or a certain class specific area. These articles use methods to extract candidate regions in the images and set them as text candidates. Special features used in these methods include color features, texture features, edge features, stroke width transformation, extreme value regions, etc. The main idea of this type of algorithm is to extract candidate regions. The most commonly used methods are SWT (Stroke Width Transform) (Meetei L S, Singh T D & Bandyopadhyay S, 2019) and MSER (Maximally Stable Extremal Regions) (Xu H, Xie S & Chen F, 2020). SWT algorithm is an algorithm based on the edge detection algorithm proposed in 2010. The advantage is that the stroke feature is basically a unique feature of stabilization. MSER was proposed as a method of extracting radioactive regions in 2002 and was not be used in text detection until 2010. The graphic structure is invariant to rotation and transformation of the image, then thresholds are set to binarize the input images. The value of the pixels inside the boundary of extreme areas are mostly 1, and the value of the pixels outside the boundary of the extreme areas are mostly 0. And the maximum stable extreme value area refers to the pixels in the area, so the difference between the pixel value of inside and outside areas is obvious, and the areas will not change with the change of the threshold, the areas of some connected areas changes little as the threshold rises.

The characteristics of MSER are as followed: The affine transformation of gray image is invariant, the stability of the area supported by the same threshold range will be selected, multi-scale detection can be achieved without any smoothing, that is, both small and large structure can be detected. Algorithms based on texture and connected regions doesn’t rely on powerful GPU (Graphic Processor Unit) and lots of training time, but it can achieve very high accuracy. However, because these methods are sensitive to pixels, and there still are some noises even after preprocessing and noise reduction, those noises in images will affect the detection results seriously.

Deep learning is the most popular type of artificial intelligence method in recent years. Since deep learning was introduced by ImageNet in 2012, everyone has noticed the great potential of deep learning in computer vision. Deep learning-based methods have significantly improved the performance of text recognition and detection tasks. One of the latest trends in text detection is to use CNN (Convolutional Neural Networks) to learn feature maps at different levels to achieve the final detection. The most commonly used methods in the field of text detection are as followed: (1) Text detection based on candidate boxes (Proposal-based), (2) Text detection based on segmentation (Segmentation-based), (3) Text detection based on a mixture of the two methods (Hybrid-based). In recent years, researchers have tried to use CNN for text detection, the most popular algorithm are as followed: Proposal-based method such as CSE (Liu Z, Lin G, & Yang S, 2019), ATRR (Wang X, & Jiang Y, 2019) and TextSnake (Long S, Ruan J, & Zhang W, 2018). Segmentation-based method such as PAN (Wang W, Xie E, & Zang Y, 2019), CRAFT (Baek Y, & Lee B, 2019) and PSEnet (Li X, Wang W, & Hou W, 2019). Hybrid-based method such as LOMO (Zhang C, Liang B, & Huang Z, 2019), IncepText (Yang Q, & Cheng M, 2018) and Guided CNN (Yue X, Kuang Z, Zhang Z, 2018).

However, there are still some problems in historical documents text detection, such as background noise, separated characters, distortion, irregular arrangement and other factors in historical documents. The existing methods can’t achieve good detection accuracy without any optimization, so it is very important to optimize the existing methods for historical documents text detection.

Research on metadata extraction and matching of scholarly documents based on implicit semantics (Jiang C, Liu J & Ou D, 2018) inspired us to combining high-level implicit semantic information from CNN with low-level texture features from MSER to achieve better detection accuracy. This paper designs a fast and effective deep learning model based on CNN and MSER. This model is called the ATD model (Ancient Text Detection Model) and it contains the ATD network (Ancient Text Detection Net) and the MSER model which based on NMS (Non-Maximum Suppression). We use the improved FCN (Fully Convolution Network, an enhanced model based on CNN) for historical documents text detection. FCN has no limits to the pixel size of input image, so ATD model can process any size of image without adjusting parameters. Moreover, the proposed ATD network includes convolution and deconvolution and the horizontal connection between them (specific description of ATD network is located at Section 3). ATD model learns the characteristics of different characters, it can deal with the noise and irregular arrangement characteristics commonly seen in historical documents, so it can achieve high detection accuracy. For separated characters (specific description located at Section 3), this paper proposes a novel method to detect text-box for single character based on ATD network and MSER synchronously, that is, utilizing the insensitivity of CNN to noise and the sensitivity of MSER to pixels between characters, and high-level implicit semantic information from CNN with low-level texture features from MSER. Then merge the different text boxes generated by ATD network and MSER, single character detection results have been greatly improved (detailed experimental comparison located at Section 4).

The main contribution of this article is to combine the advantages of CNN and MSER, that is, the accuracy of MSER on features and the robustness of CNN on noise, and high-level implicit semantic information from CNN with low-level texture features from MSER, to establish an efficient and fast ATD model for handwritten text in historical documents. According to our review of related literature, the latest research did not use CNN and MSER to complete the task of historical document detection. This article shows the novelty by using the merge algorithm as a post-processing step to obtain a text box with higher accuracy.

In this article, the first section introduces the existing detection methods and its objects, then introduces research difficulties in historical document detection. By improving the existing method, a CNN and MSER based model is proposed. The second section, background, describes the methods used by other researchers of historical document text detection, as well as the advantages and disadvantages of the methods. The third section, namely the ATD model, describes the structure of the ATD model, as well as the advantages and disadvantages of the ATD model. The fourth section is the experiment, based on the experimental data, this paper demonstrates the advantages of our method compared with other methods. The last section is Discussion, Implication, and Conclusion. It discusses how we find the problems and how to solve them, and the significance and value of the research in the field of historical document text detection. Then introduce the advantages of this method compared with other methods, as well as the academic impact of this method in this field. Last but not least, the existing problems of this method and ideas of future work.

BACKGROUND

In this section, some recent text detection methods of ancient text are introduced.

Wang M come up with an research on Tibetan ancient books text detection (Wang M, Yong C, & Li S, 2020), it mainly introduces the most commonly used algorithms and research trends of text detection in natural scenes in the past ten years, elaborates the development process of Tibetan text detection and recognition, and describes the research carried out by many researchers according to the structural features and syllable features of Tibetan characters, and the Tibetan Ancient Literature detection experiment achieved high detection accuracy.

Gong K proposes a method for extracting the content features of ancient Chinese texts based on TF-IDF (Gong K, Zhang K, & Zhang Y, 2020). It uses natural language processing technology, and taking seven texts in Zhuangzi as an example, calculates the word frequency and inverse text frequency index, and then intelligently obtains the word frequency distribution of the text and the content feature information of different texts. This method is intended to use computer technology to assist the research of ancient books, and has achieved good results.

Chen L proposes a method for Japanese ancient text recognition by deep learning (Chen L, Bing L, & Tomiyama H, 2020). This method identifies texts in ancient books by deep learning with an ancient book “usonarubesh” chosen as a dataset to test the performance of the model. In the experiments, the layout of the text is extracted into grayscale image through ARU-Net (a neural pixel labeling machine for historical document layout analysis). At the same time the original image which contains the texts is binarized, which the texts are filled with black, while the backgrounds are filled with white. Each area of text is judged by the density of black pixels and the layouts. The cut texts are then selected as the testing dataset for the trained model of deep learning CNN, AlexNet (the training dataset is ready). Finally, the experimental results are analyzed to draw conclusions and to decide the direction of future work.

Tian X published an article called “Ancient Chinese Character Image Segmentation Based on Interval-Valued Hesitant Fuzzy Set” (Tian X, Sun T, & Qi Y, 2020), to address the low segmentation accuracy caused by the rich glyph styles of ancient Chinese characters and the complex layout of ancient Chinese books, which affects the retrieval and recognition results, an algorithm for the layout image analysis of ancient Chinese books and Chinese character image segmentation is proposed. The initial segmentation results were obtained through the projection method of the layout of ancient Chinese books, and the connected component analysis of the above results was carried out to determine the rough divided blocks of under-segmentation and over-segmentation. Considering under-segmentation of adhesive Chinese characters, the improved K-means clustering method was used to segment adhesive blocks to obtain single-character images. To address the over-segmentation of character components separation, a method based on interval-valued hesitant fuzzy set is proposed. This method analyzed the features of the connected component in the block, characterized the over-segmentation connected component. The hesitant fuzzy distances between other connected components and the standard merge evaluation interval number were calculated in sequence. The connected component with the smallest distance was preferentially merged with the over-segmentation connected component until no over-segmentation connected component remained in the block. The experimental segmentation accuracy was 89.94%.

Applications of machine learning in document digitization is proposed by Dahl C M (Dahl C M, Johansen T, & Srensen E N, 2021) to give an overview of the potential for applying machine digitization for data collection through two illustrative applications. The first demonstrates that unsupervised layout classification applied to raw scans of nurse journals can be used to construct a treatment indicator. Moreover, it allows an assessment of assignment compliance. The second application uses attention-based neural networks for handwritten text recognition in order to transcribe age and birth and death dates from a large collection of Danish death certificates. It describes each step in the digitization pipeline and provide implementation insights.

Liang X proposed a comparative study of layout analysis of Tabulated historical documents (Liang X, Cheddad A, & Liang J, 2021) to stems from an industrial need, namely, a Swedish company (Arkiv Digital AB) has scanned more than 80 million pages of Swedish historical documents from all over the country and there is a high demand to transcribe the contents into digital data. Such process starts by figuring out text location which, seen from another angle, is merely table layout analysis. In this study, the aim is to reveal the most effective solution to extract document layout Swedish handwritten historical documents that are featured by their tabular forms. In short, the experiment proves that this method is effective in document layout Swedish handwritten historical documents.

He P published an article called "A Text Detection Structure Based on Attention Relational Network" (He P, & Su S, 2021). This paper proposes a relational network text detection architecture based on the attention mechanism, which is used to select and weight different granular nodes in the relational network, and also used for the nodes in the fusion of link to further reduce the computational cost. The experimental results show that the proposed framework in this paper can further predict and learn the connection relationships in the relational graph network, and has good accuracy in some indicators.

The methods of the above researchers have been proved to be effective through experiments, but for the images of historical documents with large character defects, the detection accuracy of the missing characters is greatly reduced compared to the detection accuracy of the complete character. In the above method, the text detection task of historical documents is regarded as two types of problems, this article redefines it into three types of problems (described at section 2.1.2), and uses the combination of CNN and MSER to optimize the detection accuracy of missing characters.

TEXT DETECTION MODEL

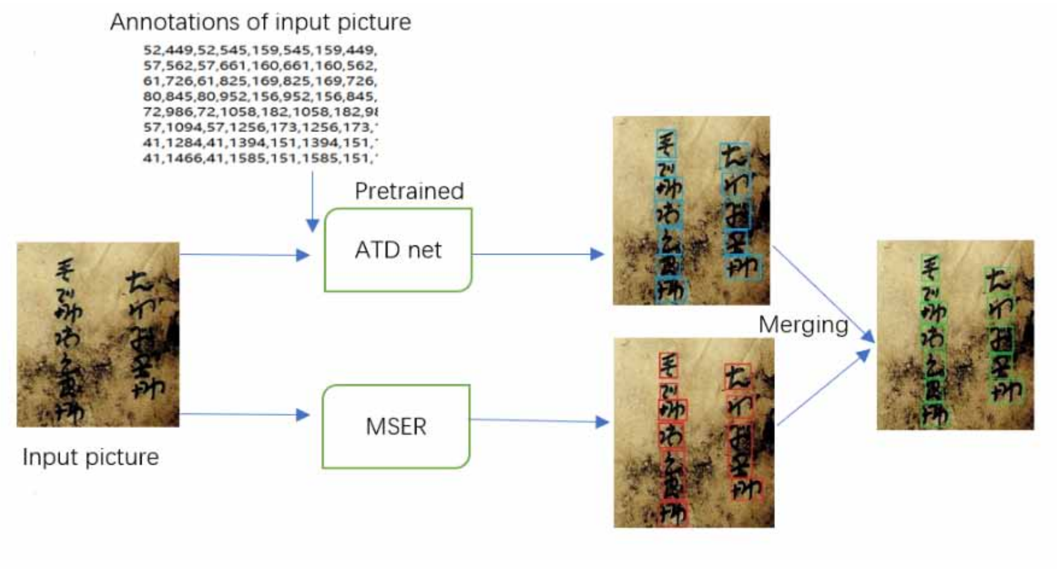
ATD model compose two parts, one is CNN-based ATD network, and the other is NMS-based MSER. The ATD network first learn the feature of historical document images with training images and the corresponding annotation information, and then classifies the characters in the test image based on the annotated data and the learned feature, then generates a candidate text-box based on the classification result. Before the image is input to MSER, the image is de-noised by non-local mean filtering, then the outline of the text is extracted using the MSER algorithm, and then the candidate box is generated from the smallest circumscribed rectangle. The two parts simultaneously output two different candidate boxes of the same character, and finally we merge the two text boxes through a merging algorithm to obtain a final text-box. Figure 1 describes the flow chart of the ATD model. The dataset used in the experiment is derived from Yi dataset (described at Sector 3.2) with its annotations, then ATD net and MSER generate two different text-box, then merge algorithm (described at Sector 3.6.1) is used to merge those two different text-box, and generate a new text-box. As shown in Figure 1 on page 20, respectively, blue text-box and red text-box represents text-box generated by ATD net and MSER, green text-box represents text-box after merging algorithm.

Classification in Ancient Text Detection

Handwritten Text Detection

Because text areas vary greatly in font, size, color, and direction, handwritten text detection is a challenging problem in image processing. Most text detection datasets have many instances of text, they either contain high contrast features compared to the background, or are blocked by other objects,

Figure 1.
ATD model flow chart



or are influenced by the background. In addition, the convolutional features of the background are very different from the text. In short, the characteristics of text and background are very different, and there is a clear boundary separating these two classes. Since CNN detector has strong convolution characteristics and strong discriminating ability, text detection becomes a binary classification problem which are text and background.

Differences in Image Detection of Historical Documents

In Yi dataset which is used in the experiment in this paper, the text often contains fuzzy, missing, weathering feature. In addition, due to the presence of noise, the contrast of the image is very low, which makes it difficult to distinguish text and background. Therefore, the CNN detector cannot learn the strong convolution features of the text. In addition, the convolutional features obtained from text areas of different densities encode the text and background information, which can confuse the detector and reduce the detection accuracy. It should be noted that the main challenge of the detection task is to solve the limited gap between the learning features of the text, the background, and the area containing part of the text.

In order to solve this problem, this paper proposes a new model. It treats detection task of ancient handwritten text as a three-category problem, as follows:

1. Text class (T): Text class refers to the complete text instance.
2. Part text class (PT): Part text class refers to texts which were covered by stains or partially missing in historical documents images or just separated characters.
3. Background class (B): B class refers to all objects in the image that do not belong to the above two categories.

Historical Documents Image Datasets

Ancient handwritten scripts not only exist in various images of historical documents, but also passed down to the world and are still being used. Therefore, our research on ancient scripts is not only technical but also important for the protection of the heritage of ancient civilization cultures.

In order to carry out comparative experiments, this article used 3 different datasets for comparative experiments. Dataset 1 is the ancient Yi dataset collected by our research team. Yi is an ancient Chinese language used by more than 4 million people in southwest region of China. It is widely used in Yunnan, Sichuan, Guizhou and Guangxi provinces. We constructed a Yi dataset which containing 470 images, includes more than 90,000 Yi characters. And 70% of them were used as training-set and 30% as test-set. Considering Yi dataset basically contains text detection difficulties such as blur, occlusion, weathering, and stains, as well as the character-level annotations, this dataset can be used to train a robust and versatile historical documents detection model.

Datasets 2 and 3 are: DIVA-HisDB dataset (Simistira F, Seuret M, & Eichenberger N, 2017), it contains medieval Latin and Italian historical manuscripts, and ANDAR-TL-1K dataset (Murdock M, Reid S, & Hamilton B, 2015), it contains 1,300 sheets from the 18th to the 19th century English document pictures of space level, two datasets are used to test the versatility and effectiveness of ATD detection model. All three datasets are shown in Figure 2 on page 20.

ATD Net Framework

In CNN, the following problems are usually observed:

(1) Feature maps in the early layers of CNN captures less semantic information.
(2) The deep layers of CNN have low-resolution feature maps, small text instances cannot be detected, while the early layer feature maps can better detect small texts.
(3) The feature map obtained after repeated sampling (down-sampling and up-sampling) makes the local accuracy of the text object very poor.

Because the image size of different historical documents datasets is not uniform, so we choose FCN as the infrastructure to construct our model. This article proposed the ATD network, which is used to detect text areas in historical document images and effectively solve the above three problems. Figure 3 on page 20 shows the structure of the ATD network, where orange, purple, light green and green blocks represents conv, maxpool, unconv, unpool, respectively.

The architecture of the ATD network consists of three different parts, as described below:

1. Convolutional layer: the convolutional layer is the front part of the network, which includes five convolutional blocks from conv 1 to conv 5, each convolutional block contains a set of continuous layers, each layer is composed of feature maps with the same resolution. Each convolution block is separated from another block by a MAXPOOL (perform maximum pooling operation) layer. The feature map resolution of each convolution block is half of the previous block, and the depth

Figure 2.
Yi dataset(left) DIVA-HisDB dataset(middle) ANDAR-TL-1K dataset(right)

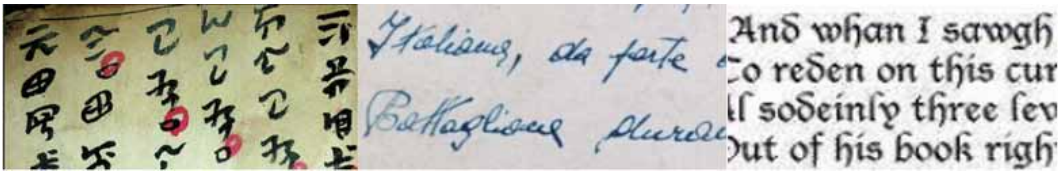
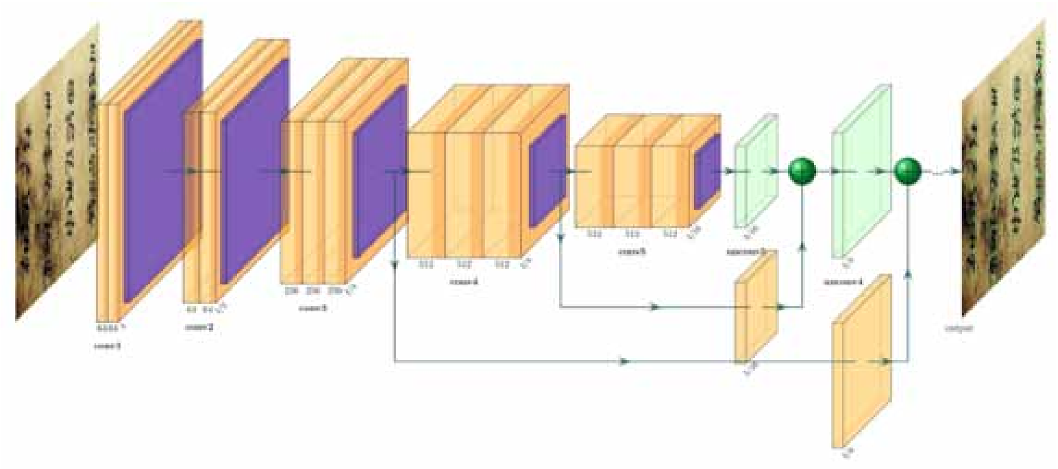


Figure 3. ATD net architecture



is twice of the previous block. Feature maps capture low-level features, while deep feature maps capture features with very strong semantic information;

2. Up-sampling layer: The up-sampling layer is used to increase the resolution of the feature map in the convolutional block after pooling. Correspondingly, the up-sampling layer also has five blocks, including five blocks from unconv 1 to unconv 5. We used UNPOOL (performs up-pooling operation, which is the inverse operation of MAXPOOL) in the middle of each block to separate it from another block, the feature map in the specific block of the specific stage of the up-sampling has the corresponding layer in the corresponding stage of the convolution which has same resolution and depth of the mapping;
3. Horizontal connection: The horizontal connection exists between the corresponding phases of the convolutional layer and the up-sampling layer. The horizontal connection is used to enhance the strong semantic information provided by the convolutional mapping and the strong semantic information in the up-sampling stem feature map. There is a separate horizontal connection between each same layer of conv and unconv, where the feature map of the last layer of the convolution stage is connected to the corresponding feature map in the up-sampling stage. In the horizontal connection, in order to reduce the result mapping depth, we used a 1×1 filter to connect these two feature maps;

The output map is obtained by convolving the final feature map with a 1×1 filter. The depth of the output map is eight channels. For the pixel P_i in the output feature map, the eight channels represent different predictions, of which the first four channels predict center coordinates, height and width of the candidate box detected by Pi, the confidence of the candidate text box of the fifth channel prediction object, and the last three channels give the confidence of the box t corresponding to three categories of T, PT and B respectively.

ATD Net Training

Loss Function

The ATD network performs three tasks simultaneously, detects candidate boxes, assigns a class-agnostic confidence score to each box, and assigns a class-specific confidence score to each box of the three classes (T, PT, and B). Based on that, this article has defined a multi-task loss function, which is defined as follows:

$$L_{mul} = L_{cls1} + »_1 L_{cls2} + »_2 L_{reg}, \tag{1}$$

Among them, L_{cls1} is the class-agnostic loss function, L_{cls2} is the total loss associated with the predicted confidence scores of the three classes, and L_{reg} is the regression loss associated with the predicted candidate box dimensions. *Pi* in the final output map predicts the coordinates, height and width of the candidate box. The candidate box is encoded as a four-dimensional vector z_i; if the predicted box has more than 0.7 overlap with the gt (ground truth) box of multiple classes, the class with the largest overlap gt box is its real label. For pixel *pi*, the true labels of the three classes are represented by □t_{ij}^{*}, where j ∈ {t, pt, b}, t_{ij}^{*} =1, provided that j is the class to which the prediction box pi belongs, otherwise $t_{ij}^{*} = 0$. Similarly, if the box at pi belongs to any class, then t_i^{*} =1, otherwise t_i^{*} =0. The definitions of various loss functions are as follows:

$$L_{cls1} = \sum_{i=1}^{res} l_{cls1}\left(t_i, t_i^{*}\right), \tag{2}$$

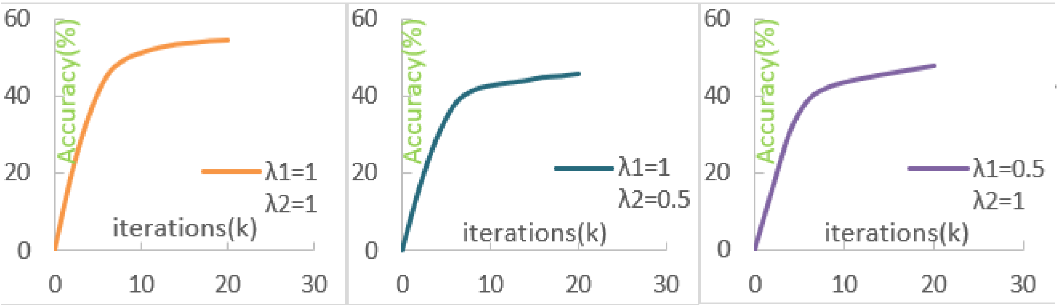
$$L_{cls2} = \sum_{i=1}^{res} \sum_{j \in \{T,PT,B\}} l_{cls2}\left(t_{ij}, t_{ij}^{*}\right), \tag{3}$$

$$L_{reg} = \sum_{i=1}^{reg} l_{reg}\left(z_i, z_i^{*}\right), \tag{4}$$

where res is the total number of pixels in the output feature map. L_{cls1} is the class-agnostic class loss associated with each detected frame. $L_{cls2\mathrm{B}}$, $L_{cls2\mathrm{PT}}$ and $L_{cls2\mathrm{T}}$ are defined for b, pt and t classes. We select l_{cls1}, $l_{cls2\mathrm{B}}$, $l_{cls2\mathrm{PT}}$, and $l_{cls2\mathrm{T}}$ as cross-entropy loss functions to facilitate training. In this paper, l_{regB}, l_{regPT} and l_{regt} are chosen as smooth l_1 loss functions because it is not very sensitive to outliers. It should be pointed out that the regression loss only considers pixels with t_i^{*} =1. Moreover, not all t-type training samples contribute to the loss function.

We accomplished experiment on loss function threshold which showed in Figure 4 on page 20. As observation in Figure 4, the horizontal ordinate means the number of iterations, and the vertical coordinates is accuracy. Where the orange line shows when the values of λ1 and λ2 are both set to 1, as each loss is equally considered, the accuracy reaches 55% at 20k iterations. And when λ1 and λ2 are set otherwise, the accuracy reached less than 48% at the same iteration time.

Figure 4.
Loss function threshold experiment, shows the different thresholds causing different accuracy



As shown in Figure 5 on page 21, the horizontal ordinates of Figure 5 are the number of iterations, the vertical coordinates are the loss and accuracy, the blue line represents the loss value, and the green line represents the verification accuracy. After experiments and observations on Yi dataset, when iterations times reaches 60k to 100k, the accuracy is approaching 100%, so we choose 60,000 as the number of iterations of the model, which consider both accuracy and time costs of training.

MSER Model

The MSER model uses non-local mean filtering to reduce the noise of the image, and then extracts the text contour using the traditional maximum stable extreme value area method, and then generates the candidate box from the smallest circumscribed rectangle. The NMS-based MSER can quickly generate candidate box.

MSER Algorithm

The idea of MSER algorithm comes from the watershed algorithm, which continuously injects water into the gully area. Some low-lying areas will be submerged by the water surface. When the water surface continues to rise, the area of these submerged areas will not change, and this unchanged area is the stable area. In the whole process of image segmentation, it is like a watershed algorithm, which can transform the image from 0 to 255. In this process, the area of some connected regions changes very little with the rise of the threshold, which is called the maximum stable extremum region.

Figure 5.
Loss function and accuracy based on iteration times

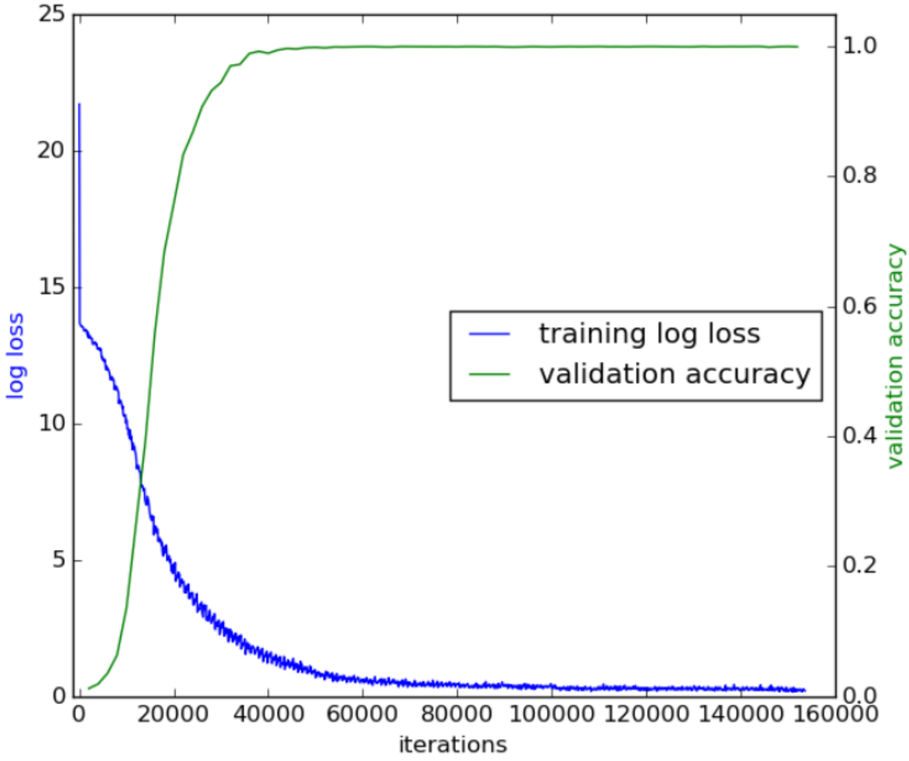


Table 1.
Algorithm for NMS

Algorithm 1

```
Input: B = {b1,...,bN}, S = {s1,...,sN}, Nt
Output: Candidate box with maximum confidence
#B is the list of initial text-boxes.
#S is corresponding detection scores.
#Nt is the NMS threshold.
begin
    D={}
    while B = ∅ do
        m=argmax S
        M=bm
        D=D∪M;B=B-M
        for bi in B do
            if iou(M, bi) ≥ Nt then
                B=B-bi; S=S-si
            end
            si=si(iou(M,bi))
        end
    end
    return D,S
end
```

NMS

NMS is an important post-processing step of the target detection model based on deep learning. In terms of target detection, the specific operation is to get the result with the highest score among all the test results, then add it to the final test result set, and then perform the other test results with this result. For comparison, if the result is that the degree of similarity (usually using IOU, that is, Intersection Over Union) is higher than the threshold, then remove it, and repeat until it is empty. NMS can easily remove most of the error candidate boxes caused by the pixels generated by the MSER. The NMS algorithm is described as follows.

Figure 6 on page 21 shows the effectiveness of the NMS with different preset IOU threshold, this test is based on Yi dataset. The horizontal ordinate is the number of iterations, and the vertical coordinates is the remaining candidate boxes after screening. As can be observed in Figure 6, if IOU value is too small, it will cause the iterations too fast and will not screening redundant candidate box. If the IOU value is too large, the iteration will be too slow, costing more time and computing resources.

In order to show the influence of IOU threshold in the detection task, we designed this experiment to test the accuracy of MSER with different IOU, as show in Figure 7 on page 22. The experiment is based on Yi dataset. When IOU is set to 0.5, MSER only reaches 23% accuracy, and ATD model reaches 58.1%. When IOU preset as 0.7, accuracy of MSER and ATD model reaches 36% and 83.4%. It's worth noticing that, when IOU=0.9, the accuracy reached 31% and 79.5% respectively. According to this test, we choose 0.7 as IOU threshold for our model.

Non-local Mean Filtering

The non-local mean filtering can be regarded as a special case of the local mean filter. The purpose is to use the area similar to the current point texture to weight the current point. That is, a weighting factor is generated based on the similarity between the weighted point and the neighborhood of the current point, namely:

Figure 6.
IOU threshold

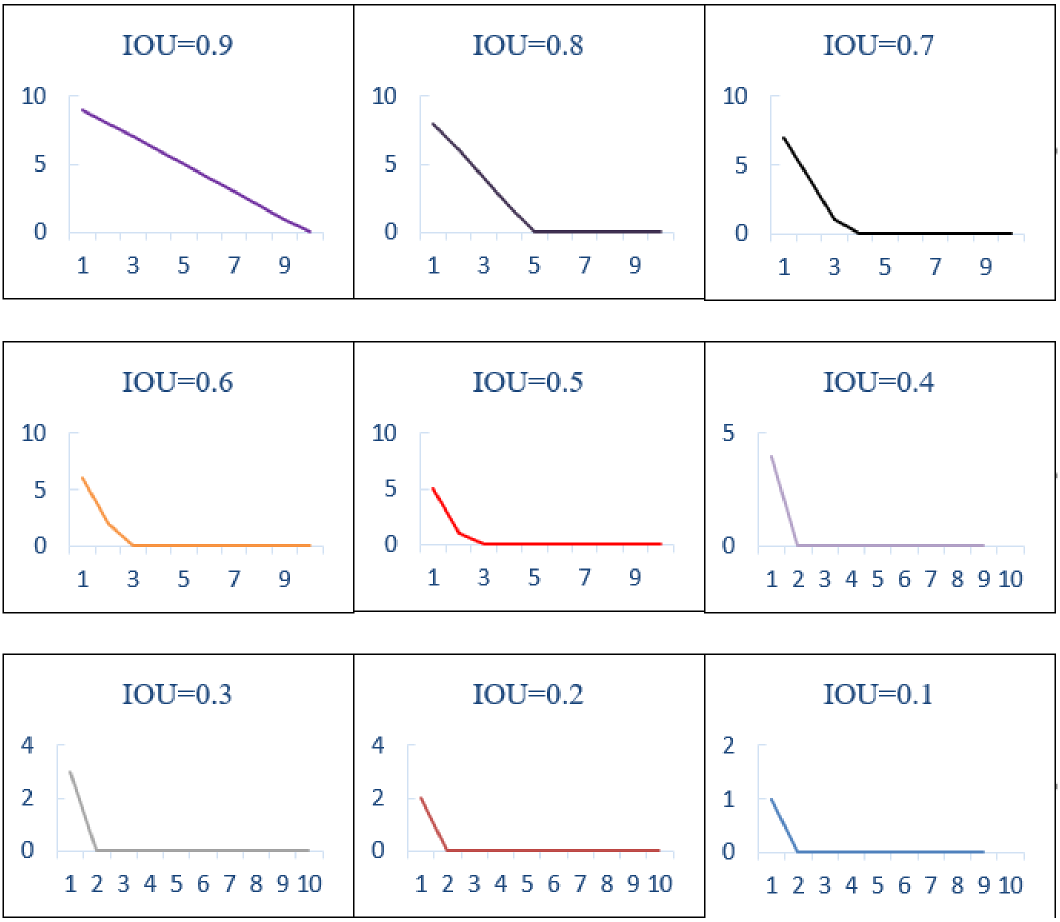
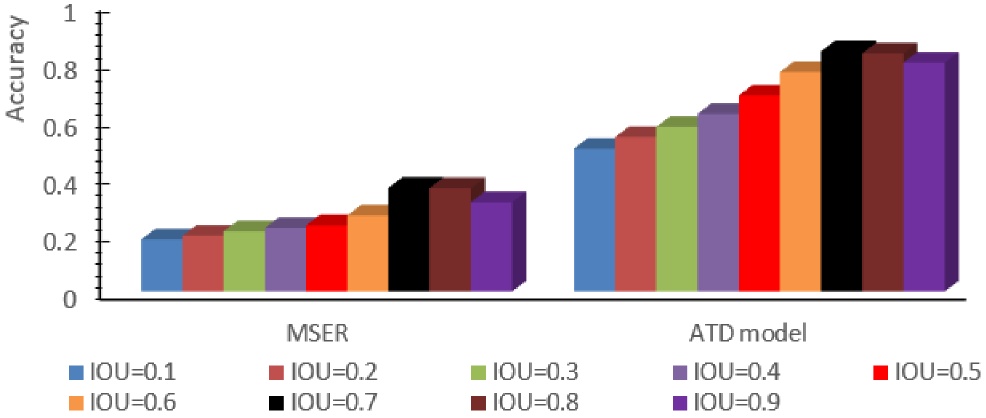


Figure 7.
Different IOU threshold cause different accuracy



$$\boldsymbol{u}\left(\boldsymbol{x}\right)=\sum_{\boldsymbol{y}\in\boldsymbol{I}}\boldsymbol{w}\left(\boldsymbol{x},\boldsymbol{y}\right)*\boldsymbol{v}\left(\boldsymbol{y}\right), \tag{5}$$

Text Box Merge

Merging

The ATD network generates a set of candidate boxes as the output of a set of characters, and its text box covers almost all non-contiguous characters. MSER can generate another set of candidate boxes for the same characters, but because the influence of historical document images, the detected candidate boxes do not contain the entire character. It is observed that in many cases, in a certain part of a character, some part is detected as a candidate box of the PT class, while the other parts of the character are detected as candidate boxes of another character. In order to solve this problem, this paper proposes a merging algorithm, which combines the T class candidate text boxes generated by the ATD network with the PT class candidate text boxes generated by the MSER model, merging algorithm are shown at Algorithm 2.

For the input image *I*, the set of candidate boxes generated by the ATD network and the MSER model are denoted by AT and MT, respectively. Algorithm 2 describes the merge process, where NEIGH(*i*) represents the set of AT-type candidate boxes in the 15×15 neighborhood of the MT-type candidate boxes (represented by *i*). The set of AT-type candidate boxes in NEIGH(*i*) is opposite to the direction in which T-type candidate boxes (represented by *j*) exist relative to MT-type candidate boxes (represented by *i*), and is given by (*i, j*). The function MAXSCORE (*i*, NEIGH(*i*)) gives the highest similarity score between the MT class candidate box (represented by *i*) and NEIGH(*i*), while MAXBOXINDEX(*i*, NEIGH(*i*)) gives the *i* a candidate box of AT class with the largest similarity.

The function check (*i*) checks whether the similarity score *i* is high enough and returns a Boolean value accordingly. According to the similarity of the MT class candidate box (represented by *i*) and the AT class candidate box (represented by *j*) in terms of their class-agnostic confidence scores, scales, aspect ratios and spatial distances between them, then the similarity is obtained. Merges two or three candidate boxes based on the number of parameters passed to it, and gives a new candidate box. Algorithm 2 is described as follows:

(1) An MT class candidate box can only search up to four candidate test boxes in its vicinity. This is because a word can be in four directions, that is, other adjacent characters may approach it.
(2) Because other parts of the same character may have been classified as MT class, the candidate text boxes formed by merging the MT class candidate text box with one or two AT class candidate text boxes are assigned to the AT class, and they will eventually be combined with this newly created AT category candidate boxes.
(3) Finally, all candidate boxes belonging to the AT and MT classes are assigned to a common FT (Final Text) class, because after the merge process is completed, all candidate boxes no longer belong to the AT or MT, but belong to the same FT class. The merge algorithm is described as follows:

It is worth noting that Algorithm 1 and Algorithm 2 are both algorithms that process text-boxes. The difference is that Algorithm 1 is used by the MSER to remove redundant text boxes formed by background noise, while Algorithm 2 is for merging text boxes generated by MSER which filtered by Algorithm 1 and text boxes generated by ATD network.

Optimization of Separate Characters by Using Merging Algorithm

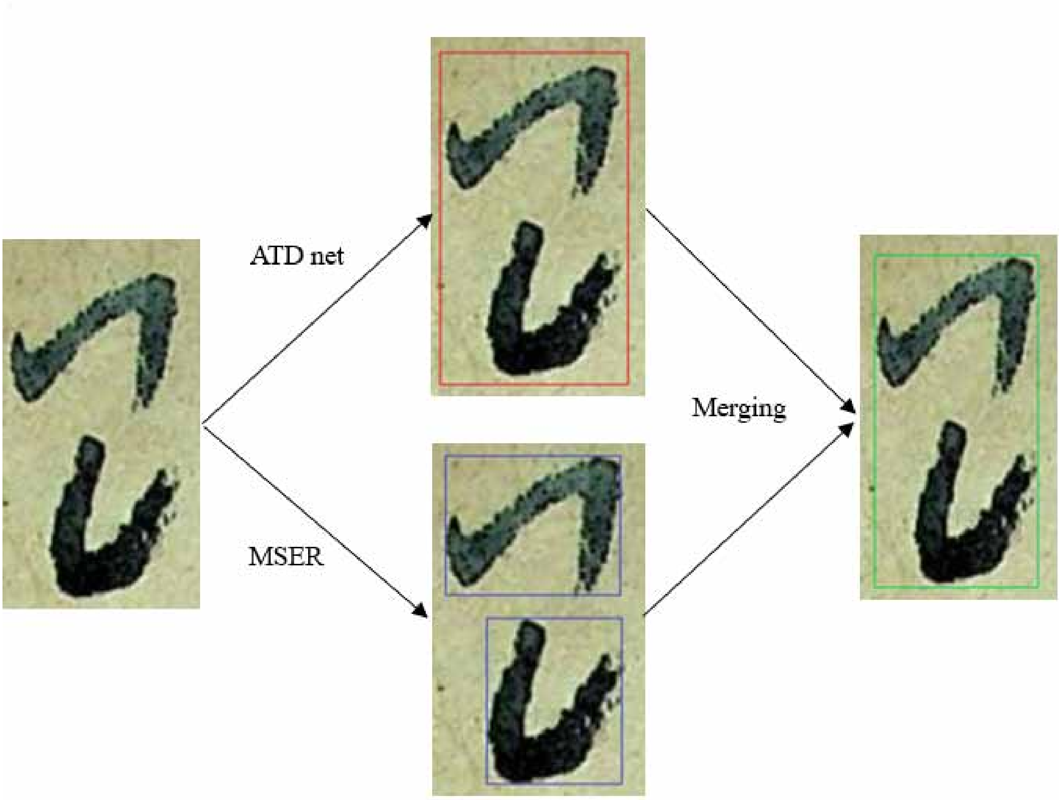
Images of historical documents often contains separated characters, that is, pixels or textures single characters are not connected to each other. An example of separated characters is shown in Figure 8 below. The influence of separated characters on the detection task is:

Table 2.
Algorithm for merging

Algorithm 2

```
Input: AT,MT
Output: FT
begin
if AT = ∅,MT = ∅,FT = ∅ do
for each i ∈ MT do
γ1i =MAXSCORE(i,NEIGH(i))
δ1i =MAXBOXINDEX(i,NEIGH(i))
if(CHECK(γ1i)==true)else
γ2i =MAXSCORE(i,OPPOSITE(δ1i))
δ2i =MAXBOXINDEX(i,OPPOSITE(i,δ1i))
if (CHECK(γ2i)==true) else
k =MERGE(i,δ1i,δ2i);
else
k =MERGE(i,δ1i);
AT = AT ∪{k}
for each i ∈ (AT ∪MT) do
MT = MT ∪{i}
end
```

Figure 8.
Merging on separate character



For algorithms based on textures or connected regions such as MSER, separate characters will cause MSER to detect two or more text boxes, causing multi-box phenomenon. For CND-based ATD networks, the separation of characters will affect the detection accuracy.

The above two methods also have different advantages and disadvantages for separate characters:

MSER can achieve a high precision for separated characters, that is, MSER is very sensitive to pixels and can generate text boxes with extremely high precision. However, it is also because of the sensitivity to pixels, separate characters will result in multiple text-boxes in a character, and pixel of noise will cause detection accuracy decline.

The ATD network can generate a complete text-box based on good learning ability and sufficient iterations of training images, but the detection accuracy of the separated characters is not high enough due to separated characters.

Based on the merge algorithm in section 3.6.1, this article quickly solved the above problem. The flow chart is shown in Figure 8 on page 22. The input image is a single-character image in ancient Yi language. After MSER processing, two text-boxes which are very close to the text texture are generated. The ATD network generated a text-box that is not precise enough. Based on the merging algorithm, the above AT class text-box and MT class text-box are merged, and the final output text-box is obtained: a complete and accurate text-box.

Advantages of Proposed Method

CNN-based ATD net uses FCN as the bone stone of architecture, so advantage of our model is that FCN-based model don’t have restriction to size of input image. FCN dose not need full connection layer, its advantages are as follows: First, it can accept any size of input image without requiring all training images and test images have the same pixel size. Second, it is more efficient because it avoids the repeated storage and convolution problems caused by the use of pixel blocks. The advantages of our model are demonstrated in Chapter 4.

EXPERIMENT

Dataset Annotations

We manually annotated through a drawing software, and annotated Yi dataset at the character level. Each character composes four coordinates totaling eight points, starting from the upper left and counterclockwise. Figure 9 on page 23 shows the Yi dataset and Yi dataset annotation respectively. DIVA-HisDB dataset and ANDAR-TL-1K dataset are collected on the internet. All three datasets are showed in Figure 9.

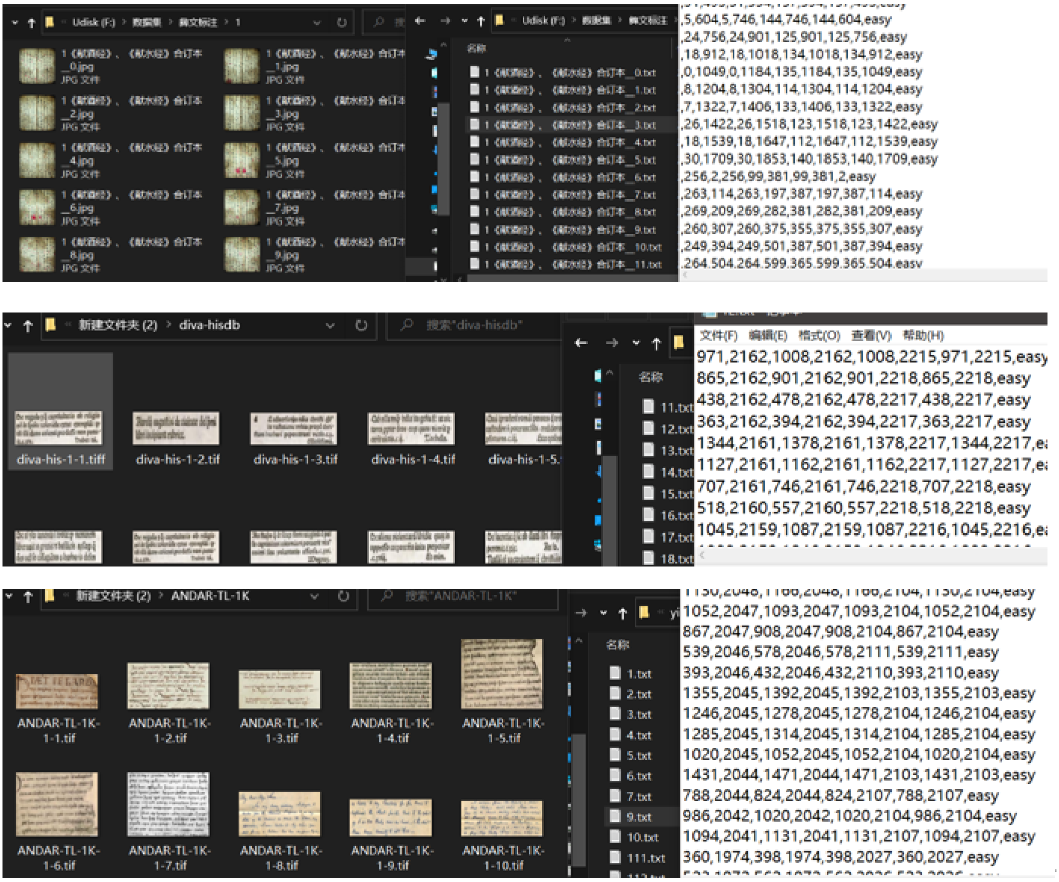
Hardware Environment

We pre-train the ATD network with Yi dataset by using NVIDIA RTX 2070s GPU, i7-9700KF CPU, 16G memory to accomplished all the experiments. We choose Ubuntu to run all tests and use python 3.7 and tensorflow 2.0 to conduct all the experiments. Because the ATD network is fully convolutional, it accepts input of all sizes, so the ATD network trained on the Yi dataset can be migrated to the detection tasks all king of historical documents. The training details of Yi data-set are as follows: 70% of the images are used as the training-set, and 30% of the images are used as the test-set.

Evaluation Index of Detection Accuracy

Recall is used to measure the proportion of the truth box G, that is, how much of the truth box G is correctly predicted. Precision is used to measure the accuracy of the detection box D, that is, how much of the predicted detection box D is true. F-measure is the weighted harmonic average of Recall and Precision. The calculation method is shown in the following formula.

Figure 9.
Data-set and annotations



```
,5,604,5,746,144,746,144,604,easy
,24,756,24,901,125,901,125,756,easy
,18,912,18,1018,134,1018,134,912,easy
,0,1049,0,1184,135,1184,135,1049,easy
,8,1204,8,1304,114,1304,114,1204,easy
,7,1322,7,1406,133,1406,133,1322,easy
,26,1422,26,1518,123,1518,123,1422,easy
,18,1539,18,1647,112,1647,112,1539,easy
,30,1709,30,1853,140,1853,140,1709,easy
,256,2,256,99,381,99,381,2,easy
,263,114,263,197,387,197,387,114,easy
,269,209,269,282,381,282,381,209,easy
,260,307,260,375,355,375,355,307,easy
,249,394,249,501,387,501,387,394,easy
,264,504,264,599,365,599,365,504,easy
```

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
971,2162,1008,2162,1008,2215,971,2215,easy
865,2162,901,2162,901,2218,865,2218,easy
438,2162,478,2162,478,2217,438,2217,easy
363,2162,394,2162,394,2217,363,2217,easy
1344,2161,1378,2161,1378,2217,1344,2217,e
1127,2161,1162,2161,1162,2217,1127,2217,e
707,2161,746,2161,746,2218,707,2218,easy
518,2160,557,2160,557,2218,518,2218,easy
1045,2159,1087,2159,1087,2216,1045,2216,e
```

```
1052,2047,1093,2047,1093,2104,1052,2104,easy
867,2047,908,2047,908,2104,867,2104,easy
539,2046,578,2046,578,2111,539,2111,easy
393,2046,432,2046,432,2110,393,2110,easy
1355,2045,1392,2045,1392,2103,1355,2103,easy
1246,2045,1278,2045,1278,2104,1246,2104,easy
1285,2045,1314,2045,1314,2104,1285,2104,easy
1020,2045,1052,2045,1052,2104,1020,2104,easy
1431,2044,1471,2044,1471,2103,1431,2103,easy
788,2044,824,2044,824,2107,788,2107,easy
986,2042,1020,2042,1020,2104,986,2104,easy
1094,2041,1131,2041,1131,2107,1094,2107,easy
360,1974,398,1974,398,2027,360,2027,easy
```

$$R\left(G_i, D_j\right) = \frac{Area\left(G_i \cap D_j\right)}{Area\left(G_i\right)} \tag{6}$$

$$P\left(G_i, D_j\right) = \frac{Area\left(G_i \cap D_j\right)}{Area\left(D_j\right)} \tag{7}$$

$$F\left(G_i, D_j\right) = \frac{2R\left(G_i, D_j\right) * P\left(G_i, D_j\right)}{R\left(G_i, D_j\right) + P\left(G_i, D_j\right)} \tag{8}$$

Text Detection Performance Comparison

This paper compares the performance of the ATD model with the most advanced text detection methods (Jiang X, Zhang X, & Xin Q, 2020; Wang, & Wenhai,2019; Yuliang Liu,Hao Chen,& Chunhua Shen,2020; Wang H, Lu P, & Zhang H,2019; Deng L, Gong Y, & Lu X, 2019; Liao M, Wan Z, & Yao C, 2019). The test results of the ATD model on Yi dataset are shown in Table 3. The table shows that the ATD model has achieved 83% Precision, 85% Recall, and 84% F-measure. In FPS (Frame Per Second), because the ATD model contains different detection methods and post-processing merge steps, the detection speed ranks second, and DBnet (Liao M, Wan Z, & Yao C, 2019) ranks first with the binary-differentiable processing speed. Training time is measured in hours, which is the time consumed by the model to learn in advance. In Yi dataset, a large number of text examples with different fonts, sizes, and directions show that the ATD model has strong robustness, and has achieved a significant improvement compared to existing methods.

In order to evaluate the effectiveness of the three classifications on the detection accuracy, to compare with the ATD model containing T, PT and B types, we constructed a model containing only T and B types, called Model-2. The training and implementation details of Model-2 are exactly the same as the ATD model, except that the number of training prediction classes is different. Table 4 shows the test results of the ATD model and Model-2 on Yi dataset. Detection accuracy of the ATD model is 15% higher than that of Model-2.

In order to evaluate the rationality of the ATD network architecture, we designed models with different layers of convolution and up-sampling layers, called Model-3, Model-4, and Model-6, the numbers represent the layers number of convolution and up-sampling layers. Table 5 shows that the detection accuracy of the ATD model (Model-5) on the Yi dataset is better than other models. As for the reason why the number of convolution layers is set to 5 is that: Too few convolutional layers will cause the learning ability of the network to be weakened under the same training time. Too many build-up layers will cause the gradient to drop, and accordingly the detection accuracy will decrease. And the experiment in Table 5 shows when convolution layers is set to 5, the precision is better than the others.

In order to evaluate the effectiveness of the merging algorithm, we used Yi dataset to test and evaluate the ATD model and the independent ATD network and MSER model, and we choose some other CNN-based method, adding MSER and merging algorithm to them to show effectiveness and superiority of the merging algorithm and our ATD net. Table 6 shows all the test results, the results show that the ATD model combining the ATD network and the MSER model has a significant accuracy improvement in the detection task. Moreover, we can easily get the comparison from Table 6, although other methods which were added with merging algorithm are better the ones without, the accuracy of our ATD net and ATD model is still in the first place.

Table 3.
Performance comparison of different methods on Yi dataset

Method	Precision	Recall	F-measure	FPS	Training time(h)
ATD model	0.83	0.85	0.84	22.5	5.3
ABAD	0.53	0.59	0.56	10.8	6.4
EAST	0.60	0.65	0.63	13.2	4.5
ABCnet	0.51	0.54	0.53	/	/
AYNIB	0.66	0.71	0.69	11	7.3
STELA	0.32	0.35	0.36	3.3	6.6
DBnet	0.72	0.77	0.75	48	8.9

Table 4.
Performance comparison between ATD model and model-2

Method	Precision	Recall	F-measure
ATD model	0.8341	0.8536	0.8439
Model-2	0.7120	0.7486	0.7303

Table 5.
Performance comparison between ATD model and models with different convolution layer

Method	Precision	Recall	F-measure
Model-5	0.8341	0.8536	0.8439
Model-3 Model-4 Model-6	0.6941 0.7694 0.7396	0.7126 0.7835 0.7411	0.7034 0.7765 0.7403

Table 6.
Performance comparison of ATD model, ATD network and MSER

Method	Precision	Recall	F-measure
ATD model	0.8343	0.8534	0.8439
ATD net	0.6772	0.7394	0.7083
MSER	0.3451	0.3609	0.3530
ABAD + Merging	0.7430	0.7617	0.7523
EAST + Merging	0.7611	0.7836	0.7720
ABCnet + Merging	0.7269	0.7441	0.7355
AYNIB + Merging	0.7983	0.8016	0.8019
STELA + Merging	0.5184	0.5378	0.5281
DBnet + Merging	0.8027	0.8236	0.8238

Because the ROC (Receiver Operating Characteristic Curve) has a good characteristic: When the distribution of positive and negative samples in the test-set is transformed, ROC can remain unchanged. In the actual dataset, sample class imbalance often appears, that is, the difference between the positive and negative samples is large, and the positive and negative samples in the test-data may change with time. The test of the effectiveness of ROC and AUC (Area Under the Curve) in Yi dataset experiment is shown in Figure 10 on page 23. As a numerical value, AUC can directly evaluate the quality of the classifier, and the larger the value, the better. And number of AUC is 0.82 shows our three types classifier is highly efficient in detection task.

In order to test the applicability of the ATD model of different historical documents dataset, we tested it on two additional handwritten historical documents datasets. It is worth noting that the DIVA-HisDB dataset and the ANDAR-TL-1K dataset are with word level annotations. While Yi dataset has character level. In order to solve this problem, we have manually adjusted the ATD model, that is, the text-boxes of several character levels output by the ATD model are merged into a word-level text box using merge algorithm, so that it can be migrated to the word-level detection task. The experimental test results are shown in Tables 7 and 8, and the visual test results are shown in Figure 11 on page 24. The green, red, blue, and black lines of text boxes in the figure represent the ATD model, ABCnet, AYNIB and DBnet. As a result of the detection, it can be clearly observed that the text-boxes of the ATD model which is the green line are closer to the text area, meaning ATD model has better detection accuracy.

Visualization of Test Results

We tested ATD model on test-image of Yi dataset. The visualization results are shown in Figure 12 on page 24. The image is composed of four kinds of ancient Yi documents with different scrolls. It can be seen that our method has high detection accuracy.

Figure 10.
ROC and AUC on Yi database detection

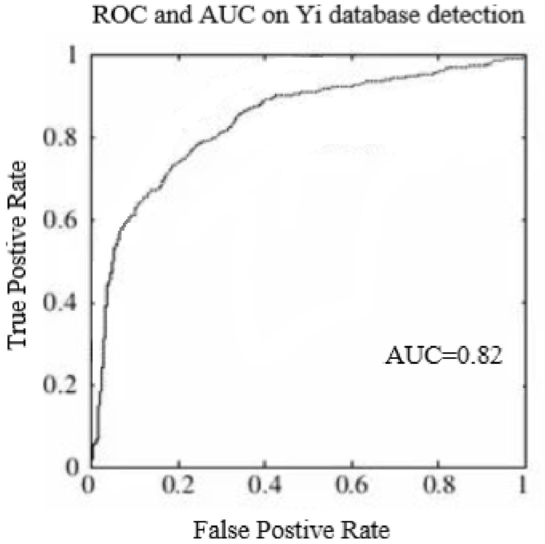


Figure 11.
Test results on DIVA-HisDB dataset(left) and ANDAR-TL-1K dataset(right)

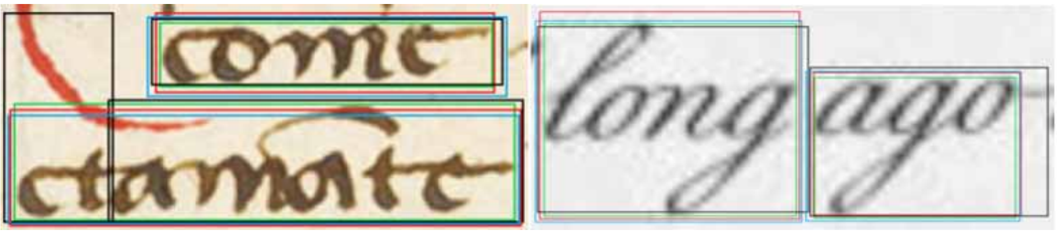


Table 7
Performance comparison of DIVA-Hisdb dataset

Method	Precision	Recall	F-measure
ATD model	0.85	0.87	0.86
ABCnet	0.72	0.77	0.75
AYNIB	0.77	0.81	0.79
DBnet	0.83	0.88	0.85

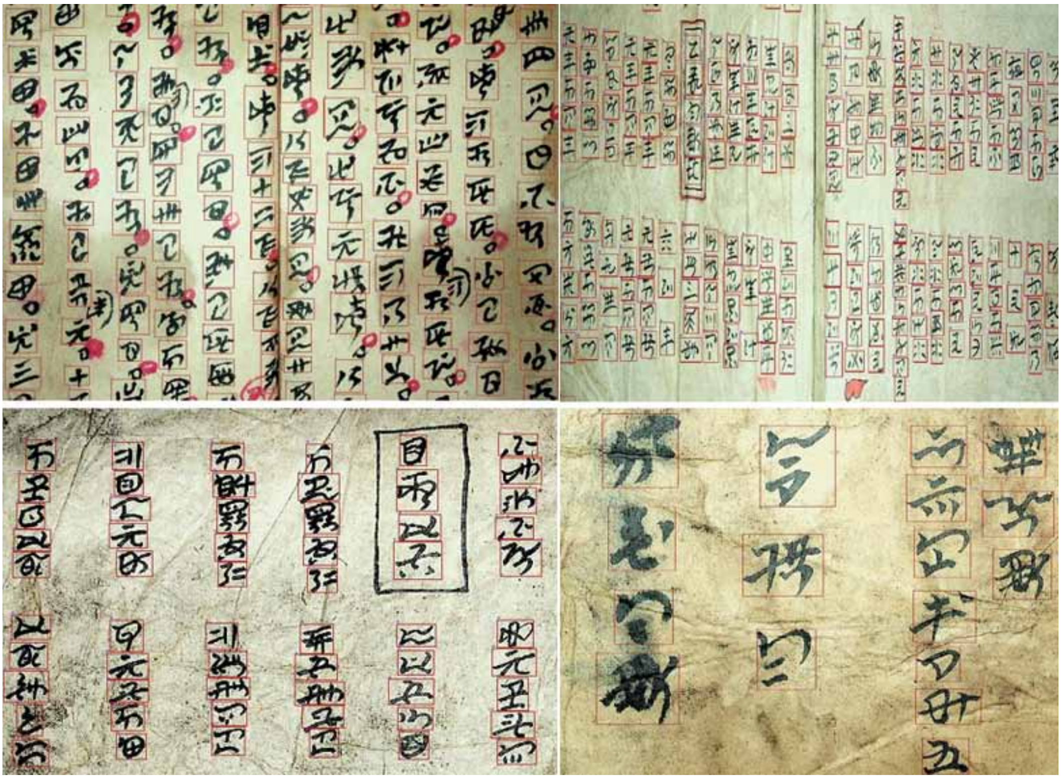
DISCUSSION, IMPLICATION, AND CONCLUSION

In this paper, starting from the binary problem of traditional handwritten text detection, the historical documents text detection problem is re-divided into three types of problems. Based on this, the ATD model is constructed. The model can achieve high-precision and high-speed detection results by using each layer of features. ATD network and MSER model simultaneously output two different candidate text-boxes from same input image. Using the merging algorithm as a post-processing step, the two

Table 8.
Performance comparison of ANDAR-TL-1K dataset

Method	Precision	Recall	F-measure
ATD model	0.79	0.87	0.83
ABCnet	0.65	0.68	0.67
AYNIB	0.73	0.77	0.75
DBnet	0.78	0.86	0.83

Figure 12.
Test results on test-image of Yi dataset



different candidate text-boxes generated by ATD network and MSER model are merged, and a better accuracy is achieved, tests on datasets of different historical documents show that our method has high detection accuracy in text detection of historical documents.

We find that there is lots of separated characters in lots of historical document such as Yi. While traditional methods and deep learning models have no good solutions for the separated characters by our tests. So we combine traditional method with deep learning method, because the traditional method has high sensitivity to the pixels of the inner character spacing and the deep learning method have learning ability of the outer character spacing, then a post-processing method is proposed to merge text-boxes. The experimental results show that our method has the best accuracy not only in Yi language, but also in English, Latin and other databases.

In the subsequent optimization of ATD model, for the word-level dataset, we plan to integrate a merge module inside ATD model, that is, merge different characters level text-box into a complete word level text-box.

In addition, for the phenomenon of mixed pictures and texts in some historical documents, we plan to use layout analysis in the ATD model to automatically classify text and images, eliminate influence caused by images in historical documents, and achieve higher detection accuracy.

ACKNOWLEDGMENT

This work is supported by the National Social Science Fund of China(19BYY171), the Open Projects Program of National Laboratory of Pattern Recognition, China Postdoctoral Science Foundation (2015M580765), and Chongqing Postdoctoral Science Foundation (Xm2016041), the Fundamental Research Funds for the Central Universities, China (XDJK2018B020), Chongqing Natural Science Foundation(cstc2019jcyj-msxmX0130), Chongqing City science and technology education research projects (KJQN201801901).

REFERENCES

Baek, Y., & Lee, B. (2019). *Character Region Awareness for Text Detection. IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Chen, L., Bing, L., & Tomiyama, H. (2020). A Method of Japanese Ancient Text Recognition by Deep Learning. *Procedia Computer Science*, *174*, 276–279. doi:10.1016/j.procs.2020.06.084

Chen, S., Wang, X., & Wang, M. (2019). A recognition method of ancient Yi language based on deep learning. *Journal of Zhejiang University*, *46*(3), 261–269.

Chen, S., Xu, H., & Mo, B. (2017). A Classification Method of Oracle Materials Based on Local Convolutional Neural Network Framework. *IEEE Computer Graphics and Applications*, *40*(3), 32–44. doi:10.1109/MCG.2020.2973109 PMID:32086199

Dahl C M, Johansen T, & Srensen E N. (2021). *Applications of Machine Learning in Document Digitisation*.

Deng, L., Gong, Y., & Lu, X. (2020). STELA: A Real-Time Scene Text Detector with Learned Anchor. *IEEE Access, PP*, (99), 1–1.

Gong K, Zhang K, & Zhang Y. (2020). Method of Extracting the Content Features of Ancient Chinese Texts Based on TF-IDF. *Electronic Technology and Software Engineering*. *163*(17), 146-147.

He, P., & Su, S. (2021). A Text Detection Structure Based on Attention Relational Network. *Journal of Physics*, *1815*(1), 12–35.

Jiang, C., Liu, J., Ou, D., Wang, Y., & Yu, L. (2018). Implicit Semantics Based Metadata Extraction and Matching of Scholarly Documents. *Journal of Database Management*, *29*(2), 1–22. doi:10.4018/JDM.2018040101

Jiang, X., Zhang, X., Xin, Q., Xi, X., & Zhang, P. (2021). Arbitrary-Shaped Building Boundary-Aware Detection with Pixel Aggregation Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *2*(99), 1–10. doi:10.1109/JSTARS.2020.3017934

Li, X., Wang, W., & Hou, W. (2018). Shape Robust Text Detection with Progressive Scale Expansion Network. *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Liang X, Cheddad A, & Liang J. (2021). Comparative Study of Layout Analysis of Tabulated Historical Documents. *Big Data Research, 24*.

Liao M, Wan Z, & Yao C. (2020). *Real-time Scene Text Detection with Differentiable Binarization*. the Association for the Advance of Artificial Intelligence, (pp. 54-63). AAAI.

Liu, Y., Chen, H., & Shen, C. (2021). ABCNet: Real-Time Scene Text Spotting with Adaptive Bezier-Curve Network. *IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 168-181). IEEE.

Liu, Z., Lin, G., & Yang, S. (2019). Towards Robust Curve Text Detection with Conditional Spatial Expansion. *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. doi:10.1109/CVPR.2019.00744

Long, S., Ruan, J., & Zhang, W. (2018). TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. *European Conference on Computer Vision*. IEEE. doi:10.1007/978-3-030-01216-8_2

Meetei, L. S., Singh, T. D., & Bandyopadhyay, S. (2019). *Extraction and Identification of Manipuri and Mizo Texts from Scene and Document Images*. International Conference on Pattern Recognition and Machine Intelligence. Springer. doi:10.1007/978-3-030-34869-4_44

Murdock, M., Reid, S., & Hamilton, B. (2015). ICDAR 2015 competition on text line detection in historical documents. *International Conference on Document Analysis and Recognition*, (pp. 1171-1175). IEEE. doi:10.1109/ICDAR.2015.7333945

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2017). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298. doi:10.1109/TMI.2016.2528162 PMID:26886976

Simistira, F., Seuret, M., & Eichenberger, N. (2017). DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts. *International Conference on Frontiers in Handwriting Recognition*, (pp. 468-479). IEEE.

Tian, X., Sun, T., & Qi, Y. (2020). Ancient Chinese Character Image Segmentation Based on Interval-Valued Hesitant Fuzzy Set. *IEEE Access, PP*, *8*(99), 1–11. doi:10.1109/ACCESS.2020.3014219

Wang H, Lu P, & Zhang H. (2020). *All You Need Is Boundary: Toward Arbitrary-Shaped Text Spotting*. The Association for the Advance of Artificial Intelligence, (pp. 351-364). AAAI.

Wang, M., Yong, C., & Li, S. (2020). Research of Tibetan Ancient Books Text Detection. *Computer Knowledge and Technology*, *16*(10), 210–213.

Wang, W., & Wen, H. (2020). Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. *IEEE International Conference on Computer Vision*, (pp. 150–159). IEEE.

Wang, W., Xie, E., & Zang, Y. (2019*)*. Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. *IEEE International Conference on Computer Vision*. IEEE. doi:10.1109/ICCV.2019.00853

Wang, X., & Jiang, Y. (2019). Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation. *IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2019.00661

Xu, H., Xie, S., & Chen, F. (2020). *Fast MSER*. *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Yang, Q., & Cheng, M. (2018). IncepText: A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection. *International Joint Conference on Artificial Intelligence*. doi:10.24963/ijcai.2018/149

Yue, X., Kuang, Z., & Zhang, Z. (2018). *Boosting up Scene Text Detectors with Guided CNN*. *British Machine Vision Conference*. IEEE.

Zhang, C., Liang, B., & Huang, Z. (2019). Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes. *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. doi:10.1109/CVPR.2019.01080

Li Rankang, from Kaixian County, Chongqing, majored in network engineering of Xinjiang University. He is currently studying in the school of computer and information science of Southwest University. He is a graduate student of grade 2018 and his tutor is Mr. Chen shanxiong.

Chen Shanxiong, Ph.D. in computing application of Chongqing University, postdoctoral in economics and management of Southwest University, associate professor, master supervisor, visiting scholar of University of South Australia in 2014. In 2003, he graduated from the Department of physics of Southwest Normal University with a bachelor's degree in science, and in 2006, he graduated from the school of computer and information science of Southwest University with a master's degree in engineering. In July of the same year, he was selected to stay in the university to engage in scientific research and teaching (one year's teaching support in Xinjiang). From March 2016 to July 2017, he took the temporary post of vice president of Information School of Guizhou Institute of engineering and applied technology. Member of CCF and IEEE, young member of the first professional committee on knowledge engineering and distributed intelligence. He has presided over and participated in more than ten projects, including National Natural Science Foundation of China, China Postdoctoral foundation, key projects of Chongqing Natural Science Foundation, 863 projects, etc. More than 30 SCI / EI retrieval papers have been published in Journal of computer science, Journal of communication, Journal of automation, Journal of information and computational science, International Journal of antenna and propagation, etc. He has been the communication reviewer of NSFC since 2014. As reviewer of more than ten journals such as journal of computer science, Journal of network and computer applications, Journal of communication, etc., mainly engaged in data mining and pattern recognition

Zhao Fujia, born in Sheyang County, Yancheng City, Jiangsu Province, studied in Yancheng Institute of technology, majoring in software engineering. He is currently studying in the school of computer and information science of Southwest University. He is a graduate student of grade 2018 and his tutor is Mr. Chen shanxiong.

Qiu Xiaogang, from Guangyuan, Sichuan Province, is a graduate student of 2018 in the school of computer and information science of Southwest University, majoring in computer system structure, and his tutor is Chen shanxiong.