Form location and extraction based on deep learning

Zhen Zhang¹, Dengyin Zhang¹2*, Tianyu Jin¹, Min Zhang¹

¹Jiangsu Key Laboratory of Broadband Wireless Communication and Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China
²School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China

*Corresponding author e-mail: zhangdy@njupt.edu.cn

Abstract. To obtain users’ information on form images with different types and resolutions, a form location and extraction method based on deep learning is proposed in this paper. We adopt two classification models to better the detection accuracy: model A is applied to detect the smallest rectangular area containing the main information; model B recognizes users’ names, phone numbers and addresses. Especially, the proposed scheme with the certain anchor sizes and positive sample selecting method achieves 82.44% accuracy on the challenging dataset from reality.

1. Introduction

As an important information carrier, forms have been widely utilized in real life and work. Some specific patterns and symbols in forms are likely to contain essential information that the user is interested in, such as the order number, the specific item of the invoice, the amount, the address and mobile phone number in the express waybill. The method of manually inputting data is time-consuming and laborious, and it is easy to make mistakes. Consequently, the automatic extraction of form images has a strong requirement, which can greatly reduce the workload and improve the efficiency.

The chief process of form automation processing includes the acquisition of waybill images, the extraction of the crucial information, the recognition [1]. Pre-acquiring key areas of forms is more convenient and accurate to identify the content in forms. This paper selects the express waybills with the most complicated background as the research object, which has become challenging due to varying illumination conditions, complicated scenes, image distortions and rotations caused by manual shooting. The principal job is to extract the text areas related to the user information in the waybill pictures, such as the names, numbers and addresses. The output of this procedure, which has broad prospects, could be employed for subsequent text recognition, also for establishing character image databases.

Lu et al. [2] offered a logo matching approach to extract the consequential data of forms. The main points that are scale-invariant were extracted from the logo pictures and waybill photos. Then they searched these points by random sample consensus in other photographs. Once the mapping matrix is calculated, the information would be obtained. Sachdeva et al. [3] utilized a blank form template to align with the form to be matched. And based on the statistical features of the image, the morphological method was applied in the work to remove the frame line of the image and separate the areas. An efficient approach was proposed by Ashutosh et al. [4] who analyzed the constituent elements of the form, such as icons, command fields, line segments, etc., and described them in a...
specific language. When identifying different types of forms, manual descriptions require to add. These schemes are sensitive to the format and complex scenarios accordingly.

Notably, users’ information could be seen as assorted targets. Then the form extraction could be transformed into the object detection, which not only can locate areas of users’ information, also classify it.

Recently, Deep Convolutional Neural Networks have been more and more popular, and more and more object detection methods with high precision have been proposed, for instance, the SSD [6], the YOLOs [7, 8], and the Faster R-CNN [5]. The SSD and the YOLOs are both regression based algorithms, the difference is that the SSD employs an additional layer to predict the default box offsets to improve the accuracy. Unlike them, the R-CNNs are prediction based approaches. A network called the Region Proposal Network (RPN) is trained to generate candidate boxes, which guarantees a very high accuracy.

The Faster R-CNN has achieved promising results for the classification tasks in quite a few fields (e.g., [9]), but there is no direct application on waybill detection. Therefore, we aim to employ the Faster R-CNN to locate and extract the valuable information in the waybills.

2. Form location with the Faster R-CNN

2.1. The Architecture of the Faster R-CNN

As illustrated in the right part of Fig. 1, the Faster R-CNN contains two core modules: 1) the RPN, which mainly generates region proposals from future maps; and 2) the Fast R-CNN, which not only is responsible for classifying the proposals, and refining the bounding boxes.

![Figure 1. The structure of the Faster R-CNN.](image)

With the alternating optimization, both the RPN and the Fast R-CNN could be trained to share the feature maps pulled from the last convolutional layer (e.g., “Conv5_3” for VGG16 [10]). In this way, computational complexity significantly could be reduced.

2.2. The Proposed Form Detection Method

The results of applying the Faster R-CNN directly to detect the users’ information are not satisfying due to the extra boxes, error boxes, and missed boxes.

To approach the above problems, we show a framework in Fig. 2, which comprised of two Faster R-CNN models to reduce the interference of shadows in the background. The model A is employed to identify the smallest rectangular area (called area M) containing the main information, and the model B recognizes recipients’ and senders’ names, phone numbers and addresses.
To improve the detection accuracy, two more strategies are proposed as follows:

The sizes of the anchors of RPNs is calculated by clustering the sizes of the ground-truth boxes. So, we employ 3 scales of \{64, 128, 256\} and 3 aspect ratios of \{0.3, 0.5, 0.8\} for anchors of model A and 3 scales of \{32, 64, 128\} and 3 aspect ratios of \{0.2, 0.5, 1\} for model B.

Besides, we propose a positive sample selecting method to balance the number of the positive and negative sample, which while not adding too much calculation, adds many positive samples containing targets’ message. In details, we redefine the ground-truth box \(gt'_{ij}\) corresponding to the candidate box \(a_i\) (see Fig. 3). If the intersection over union (IOU) value of \(a_i\) and \(gt_j\) is greater than a certain value, \(a_i\) could be seen as a positive sample. Otherwise, it is marked as a negative sample.

Specifically, the location information of \(a_i\), \(gt_j\) and \(gt'_{ij}\) can be expressed using the following equations:

\[
a_i = (a_{ix}, a_{iy}, a_{ix}, a_{iy})(i = 0, 1, 2, \ldots) \tag{1}
\]

\[
{gt}_j = (gt_{jx}, gt_{jy}, gt_{jx}, gt_{jy})(i = 0, 1, 2, \ldots) \tag{2}
\]

\[
{gt'}_{ij} = (\max\{gt_{jx}, a_{ix}\}, \min\{gt_{jy}, a_{iy}\}, gt_{jx}, gt_{jy}) \tag{3}
\]

where \(a_{ix}, a_{iy}, a_{ix}, a_{iy}\) and \(gt_{jx}, gt_{jy}, gt_{jx}, gt_{jy}\) represent the coordinate values of the upper left and lower right corners of \(a_i\) and \(gt'_{ij}\) respectively.
Then, the $\text{IOU}(a_i, \text{gt}_y)$ can be defined as:

$$\text{IOU}(a_i, \text{gt}_y) = \frac{\text{area}(a_i) \cap \text{area}(\text{gt}_y)}{\text{area}(a_i \cup \text{gt}_y)}$$

(4)

where $\text{area}(\ast)$ means the acreage of $(\ast)$.  

3. Experiments Results and Analysis

3.1. Experiments Setup and Dataset

The development environment is built on the deep learning tools of Matlab2018b. The main computer specifications executing all the experiments are shown in Table 1.

| Equipment             | Configuration                      |
|-----------------------|------------------------------------|
| CPU                   | Intel Xeon E5-1620 v4 @ 3.50GHz    |
| Graphics Card         | NVIDIA TITAN Xp, 12 GB GDDR5       |
| Memory                | 64 GB DDR4                         |
| Operating system      | Ubuntu 16.04.5                     |

The waybill images in this article are all from reality, shooting by the staff. Since the pixels of the captured images are too high, all the pictures are subjected to compression processing and gradation processing. After pre-processing, 1952 pictures are obtained, 1652 of which are for training and 300 of which test the model.

To comprehensively evaluate the performance of our framework, the Faster R-CNN and the YOLO v3 in deep learning are included for comparison.

3.2. Performance Evaluation

We choose the Average Overlap (AO) and mean Average Precision (mAP) to characterize the performance of the model. The AO indicates the degree of the overlap between the detection boxes and the ground-truth boxes. The mAP stands for the overall situation of achieving correct location across the test data. The AO could be defined as the IOU value of the detection boxes and the ground-truth boxes:

$$AO = \frac{1}{nl} \sum_{i=1}^{nl} \text{IOU}(\text{gt}_i, v_i)$$

(5)

where $nl$ is the total number of object areas, $\text{gt}_i$ represents the ground-truth box and $v_i$ is the detection box.

The mAP could be expressed as:

$$\text{mAP} = \frac{\text{num}(AO \geq T)}{N}$$

(6)

where $T$ is the threshold that determines the predicted box is correct or false, $\text{num}(AO \geq T)$ is the number of the positive detected boxes when the value of the AO is greater than or equal to $T$, and $N$ is the number of the test set.

3.3. Results and Analysis

The values of the mAO and mAP are calculated when the threshold values $T = \{0.5, 0.6, 0.7, 0.8, 0.9\}$, presented in Table 2.
Table 2. The values of the mAP and the mAO.

| Method     | mAP   | mAO   |
|------------|-------|-------|
|            | 0.5   | 0.6   | 0.7   | 0.8   | 0.9   |
| Faster R-CNN | 0.9329 | 0.9081 | 0.8690 | 0.8127 | 0.7834 | 0.8076 |
| YOLO v3     | 0.8633 | 0.8174 | 0.7828 | 0.7389 | 0.6899 | 0.7123 |
| Our method  | 0.9577 | 0.9312 | 0.9031 | **0.8683** | 0.8090 | **0.8244** |

From table 2, as \( T \) increases from 0.5 to 0.9, the mAPs of the three schemes decrease, which can be attributed to the fact that the content of the waybill images has a large degree of freedom, and there are many characters or shadows similar to the key information. Also, it can be seen that the proposed approach has a mAO of 0.8244, indicating that the bounding boxes predicted by the model and the ground-truth boxes have an average overlap rate of more than 80% and the predicted boxes have been substantially overlapped with the ground-truth boxes, which means our method has good performance. The YOLO v3 based one stage detection exchanges speed for accuracy, accordingly it has the smaller mAP and mAO.

Furthermore, we choose \( T=0.8 \) as the confidence of the algorithm. The accuracy of the objects including area M, name, phone number and address from the three methods are shown in Table 3.

Table 3. The accuracy of the objects detection.

| Method     | Area M | Name | Phone number | Address |
|------------|--------|------|--------------|---------|
| Faster R-CNN | 0.9540 | 0.8846 | 0.9028       | 0.7602  |
| YOLO v3     | 0.9182 | 0.8491 | 0.8174       | 0.7087  |
| Our method  | 0.9589 | 0.8912 | 0.9265       | 0.8189  |

Note that the phone numbers’ recognition accuracy is higher than the names’ and addresses’ because of the fixed length. Some users are merchants, the length of whose name is similar to some addresses. Some of the addresses are shorter as the name boxes, which affects the accuracy of the model. As for the detection of the area M, the reason why the disparity between the Faster R-CNN and our method is not manifest is that the area M belongs to the large targets, which the offered strategies by this paper are not aimed at. With respect to the detection of the valuable information including name, phone number and address, the accuracy of the proposed method is higher than the Faster R-CNN and YOLO v3, which benefits from the strategies proposed in 2.2. Comparing with the Faster R-CNN, the proposed scheme increases the accuracy of the elongated address boxes by 5 percentage points.

4. Conclusion

In this paper, a form location and extraction method based on the improved Faster R-CNN is proposed to locate and extract the key information. The experimental results show the proposed approach with cascade models achieves high accuracy in waybill images with complex background. In particular, the high performance of the system is guaranteed by the proposed anchor sizes and the positive sample selecting method.

About the future work, we are willing to further improve the speed and the accuracy of the algorithm by creating heterogeneous models that combining the power of the DCNN and traditional computer vision methods.

5. Acknowledgments

This work was financially supported by National Natural Science Foundation of China (No.61571241 and 61872423), Industry Prospective Primary Research & Development Plan of Jiangsu Province (No.BE2017111), the Scientific Research Foundation of the Higher Education Institutions of Jiangsu Province (No. 19KJA180006), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX18_0912).
References

[1] Sharma D V, Lehal G S. Form field frame boundary removal for form processing system in Gurmukhi script. 2009 10th International Conference on Document Analysis and Recognition. IEEE, 2009: 256-260.

[2] Lu S, Zhao J, Lu Y. Express waybill extraction on parcel images by logo matching. 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016: 2861-2865.

[3] Sachdeva R, Sharma D V. Data extraction from hand-filled form using form template. International journal on recent and innovation trends in computing and communication, 2015, 3(8): 5311-5317.

[4] Mishra A, Mondal P, Banerjee S. 2D/3D Non-rigid Image Registration by an Efficient Demons Approach. 2014 IEEE 27th International Symposium on Computer-Based Medical Systems. IEEE, 2014: 481-482.

[5] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 2015: 91-99.

[6] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. European conference on computer vision. Springer, Cham, 2016: 21-37.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[8] Redmon J, Farhadi A. Yolov3: An incremental improvement. 2018, https://arxiv.org/abs/1804.02767.

[9] Pham M T, Lefèvre S. Buried object detection from B-scan ground penetrating radar data using Faster-RCNN. IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2018: 6804-6807.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, https://arxiv.org/abs/1409.1556.