Form Recognition Based on Lightweight U-Net and Tesseract after Multi-level Retraining

Gang. Li
School of Mechanical Engineering
Xi'an Jiaotong University
Xi'an, China
lg618181@stu.xjtu.edu.cn

Junhui. Huang*
School of Mechanical Engineering
Xi'an Jiaotong University
Xi'an, China
* huangjh@mail.xjtu.edu.cn

Zhao. Wang
School of Mechanical Engineering
Xi'an Jiaotong University
Xi'an, China
wangzhao@mail.xjtu.edu.cn

Jianmin. Gao
School of Mechanical Engineering
Xi'an Jiaotong University
Xi'an, China
gjm@mail.xjtu.edu.cn

Kun. Chen
School of Mechanical Engineering
Xi'an Jiaotong University
Xi'an, China
chenkun@mail.xjtu.edu.cn

2021 International Conference of Optical Imaging and Measurement (ICOIM) | 978-1-6654-0354-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOIM52180.2021.9524385

***Abstract*—With the rapid development of Internet information technology and the advancement of enterprise digitization, the digitization of paper forms has also received extensive attention. The automatic conversion of paper form documents into electronic form documents mainly faces three problems. The first is that the format of the form file is diverse and the structure is complex. This article uses the XML file of the form to accurately analyze the structure of the file, which is more accurate than the current semantic segmentation method. The second problem is table area detection, this paper uses traditional algorithms to find the contours of the candidate table area, and screens according to the characteristics of the table area to complete the detection and extraction of the table area. The third is that the recognition of the table text is more difficult, not only the interference information such as the table frame will also affect the accuracy of text recognition, and the type of text information in the table is complex, including Chinese, English, numbers, symbols and mixed types, which bring huge challenges to text recognition. This paper uses the lightweight U-Net network model to segment the text area at pixel level, eliminating the interference information of text recognition. The neural network of Tesseract was retrained in a multiple, multi-level manner, and successfully realized the recognition of complex types of text information with an accuracy of about 96%. Based on deep learning and XML table structure analysis algorithm, this paper realizes the recognition of paper version of the form file and the reconstruction of the electronic version of the file.**

Keywords-form recognition; XML; lightweight U-Net; Tesseract; multi-level retraining;

I. INTRODUCTION

Tables are an important form of information expression. The standard organization structure is used to store information, which is convenient for information retrieval and analysis[1]. However, there are many problems with manual processing of form files, such as a large number of forms, too much manual labor, and easy data processing errors, which can bring immeasurable losses to the production of enterprises. Therefore, with the advancement of enterprise digitization, form recognition has received extensive attention.

In the field of table area detection, the current mainstream detection algorithms are divided into two types: candidate area-based algorithms and regression-based algorithms. The candidate area-based algorithms are represented by Faster R-CNN. Gilani[13] and Sun[14] applied Faster R-CNN[4] to table area detection tasks and achieved good results. Huang[5] used the regression algorithm-based YOLOv3 [6] model for table area detection, and introduced the anchor frame optimization strategy and post The processing operation greatly improves the calculation accuracy. In the field of table structure recognition, most of the current mainstream methods are based on deep learning methods such as semantic segmentation models, target detection, and graph neural networks. Siddiqui[7] successfully identified the rows and columns of the table using semantic segmentation technology based on the FCN framework. Siddiqui et al. [8] regard the recognition of rows and columns in the table structure as a target detection problem, documents are regarded as scenes, rows and columns can be regarded as objects, and they also have achieved good results; Qasim[9] Describe the table structure recognition problem as a graph problem compatible with graph neural network, and try to use graph neural network to solve this problem. Of course, there are also end-to-end table detection and structure recognition algorithms. Prasad[10] proposed an end-to-end convolutional neural network model CascadeTabNet based on deep learning which can achieve pixel-level segmentation of

978-1-6654-0354-2/21/$31.00 ©2021 IEEE

each table image For table instances and cell instances, in order to achieve high-precision instance segmentation, Cascade RCNN [11] and HRNet [12] are selected as the main body of the network. In the field of text recognition, the algorithm is mainly divided into two steps, feature extraction and sequence conversion. Feature extraction usually uses convolutional neural networks such as VGG[13], ResNet[14], DenseNet[15], etc. Sequence conversion commonly used CTC[16] and Seq2Seq[17] and other sequence transformation models. Among them, the Convolutional Recurrent Neural Network (CRNN) model[18] is the most classic character recognition model, which can accurately recognize long text sequences.

Form recognition is to use table images to obtain table structure information and cell content information, and complete form reconstruction. But the form image contains a huge amount of information, such as table lines, cell content, etc. Moreover, the purpose of the table recognition task is very complicated, and it is necessary to obtain the logical information of the cell and the content information corresponding to the cell. It is difficult to implement table structure analysis, content recognition, and table reconstruction based on deep learning. Due to the large amount of data and complex tasks, the network model will be very complicated. The model has a huge amount of parameters and calculations, and the training and implementation of the model is very difficult. In addition, it is difficult to balance the accuracy and performance of the model and it is difficult to apply the algorithm to actual production.

Therefore, this article focuses on the application-level form recognition algorithm. Based on the traditional image contour finding algorithm to detect and extract the table area in the image. Accurately obtain the structure information of the form file through the XML-based parsing method proposed in this article. And use the exclusive data set for multi-level retraining to perform multiple and multi-level retraining on Tesseract, so that the accuracy of text recognition meets the application requirements. Finally successfully realized the identification of the paper version of the form file and the reconstruction of the electronic version of the form.

II. FORM RECOGNITION ALGORITHM

A. *Brief description of algorithm.*

The paper-based form file recognition algorithm proposed in this paper mainly has 6 steps, as shown in Figure 1. The first step is to detect the table area, extract the table area to be recognized in the image and perform tilt correction. The second step is to parse the structure information of the form, and accurately obtain the structure information of the table through the XML-based parsing method. The third step is to carry out the cell coordinate information conversion and decompose the table area, extract the independent cell image of the form. The fourth step is to detect and locate the text of the cell image, and eliminate the influence of factors such as borders on the accuracy of text recognition. The fifth step is to complete the text information of the cell image. The sixth step is the reconstruction of the electronic form. Finally, the form recognition algorithm proposed in this paper successfully completes the identification of paper form files and the reconstruction of electronic files.

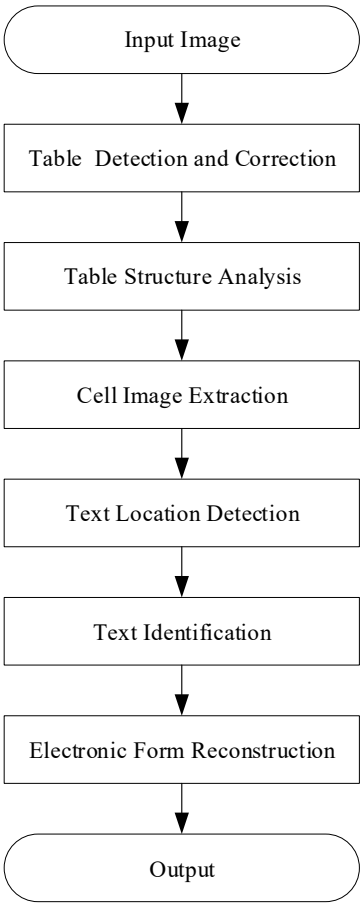

Figure 1. The main process of the form recognition algorithm

B. *Table Area Detection*

First, scan the paper version of the form file to obtain the corresponding form image data. Then first perform the form area extraction operation. First finds all the contour information in the image. In order to accurately extract the table area, the contour information needs to be further filtered. This article cleverly makes use of the difference between the table area and other contours. That is the contour edge relationship and contour area of the table area are obviously different from other contours. Therefore, the screening condition is that the opposite side of the contour edge needs to be approximately parallel, and the adjacent side needs to be approximately perpendicular. On the basis of the above contour edge condition, the condition of the maximum contour area needs to be satisfied. The outline to be filtered meets all the above conditions is the target table area. After obtaining the table area in the image, due to the phenomenon of a certain image tilt during printing and scanning of the table image, which will affect the performance of the algorithm, it is necessary to use the affine transformation algorithm to perform certain correction work on the table area. Affine transformation is an important transformation in the two-dimensional plane. It has a wide range of applications in the field of image graphics. In two-dimensional image

transformation, it generally expresses as (1). (x, y) are the coordinates of the table area before affine transformation, R is the rotation component in the transformation matrix, T is the translation scaling component in the transformation matrix, and (X, Y) are the coordinates of the table area after affine transformation.

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = \begin{bmatrix} RT \\ 01 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (1)$$

According to the target contour information obtained by the table area extraction operation, calculate the four corners of the rectangular area of the table and the length of each side of the area, and obtain the coordinate information before the affine transformation. Take one of the corners as the reference and use the side length information to calculate the coordinate information of the target table area after affine transformation is brought into the affine transformation formula to calculate the corresponding affine transformation matrix to complete the tilt correction of the table area.

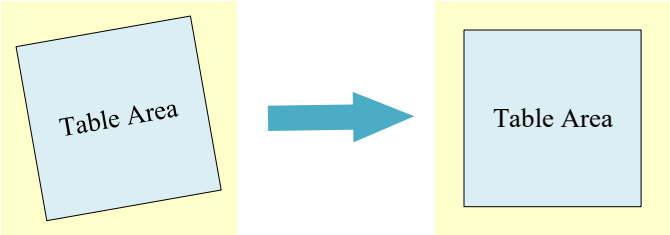

Figure 2. Table area detection and correction effect diagram

C. *Analysis of Form Structure Based on XML*

After completing the extraction of the table area of the table image and the tilt correction, the next step is to use the XML file of the table to accurately parse the corresponding structural information. XML is just an extensible markup language, a markup language used to mark electronic files to make them structured. It can be utilized to record data in form files and define data types and structure information. Combined with the general production rules of the form file, the analysis of the XML file can obtain the structure information corresponding to the paper version of the form file, and the coordinate information and length and width information of each cell in the corresponding EXCEL electronic source file can be obtained by calculation. And store the structure information obtained by analysis to prepare for the subsequent algorithm.

D. *Cell Image Extaction*

Then carry out the conversion of the cell coordinate information in the EXCEL electronic source file to the cell coordinate information in the form file. Ignoring factors such as image distortion, the coordinate system of the corrected image table has a certain scale and movement relative to the table of the EXCEL electronic source file, so the coordinate system conversion calculation is completed according to the (2). Among them, (X, Y) is the actual coordinate of the cell in the image, (x, y) is the cell coordinate information in the EXCEL electronic source file, (H, W) the length and width information of the table area in the image, (h, w) is the length and width information of the table area in the EXCEL electronic source file, (x_lu, y_lu) is the coordinate information of the upper left corner of the table area in the image.

$$(X,Y) = (x,y)\frac{(H,W)}{(h,w)} + (x\_lu, y\_lu) \quad (2)$$

According to the calculated coordinate information of the image table cell, the cell image is cropped to obtain an independent cell image, and the corresponding number is used to facilitate the subsequent table reconstruction.

E. *Text Location Detection Based on Lightweight U-Net*

After obtaining an independent cell image, it is necessary to detect and extract the text area to avoid the influence of the cell border on the accuracy of text recognition. In this paper, based on U-Net[20] network, the cell image is segmented semantically, and the text area is extracted. U-Net is composed of an encoding-decoding architecture. The encoding structure completes the compression and feature extraction of the input image through multiple downsampling. In the decoding structure, the model completes the upsampling of the underlying features through the upper convolutional layer, and finally restores to the original image scale size. At the same time, the down-sampling feature layer and the up-sampling feature layer of the corresponding scale are connected through the feature fusion channel, and multi-scale feature fusion is realized through the stacking method, making full use of multi-level semantic information. Finally, the pixel-level segmentation of the input image is completed, and the end-to-end image segmentation technology is realized.

The convolutional layer in the U-Net model architecture is relatively large, and the training parameters are more computationally expensive. Taking into account the performance of the model and the difficulty of the algorithm, this paper uses a deep separable convolution module to replace the ordinary convolution layer to achieve lightweight operations on the U-Net model, which greatly reduces the training parameters and the amount of calculation. At the same time, in the process of making the model training data set, it is necessary to consider the distribution of the actual table data, such as the shape and size of the cell, the type of characters in the cell and the existence of blank cells, so the training sample needs to be balanced according to the actual distribution to ensure the accuracy and robustness of the model's predictions. The final cell text area extraction effect is as figure 3.

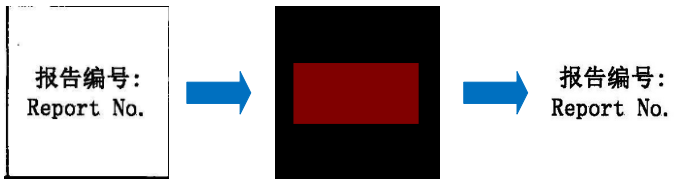

Figure 3. Cell text location detection

F. Text Identification

To recognize the text area image extracted from the cell, this paper is built on Tesseract and retrains the network to complete the text recognition work. Tesseract[22] is an open source OCR engine developed by HP Labs and maintained by Google. It is open source, free and supports multi-language and multi-platform deployment. However, Tesseract is designed for English recognition, and the characters in the table for this article are a mixture of multiple types of languages, including printed Chinese, English, numbers, symbols, and their random mixture, which is difficult to recognize. Therefore, in order to ensure the recognition accuracy requirements of the form characters, this paper adopts the method of multi-level retraining to gradually retrain Tesseract, and gradually improve the text recognition ability of Tesseract. Tesseract is gradually retrained several times in the order of English, English and number mixed type, English, number and symbol mixed recognition, Chinese and multiple types randomly mixed. In the process of making the training data set, first collect table file data, analyze the data distribution of various texts, and then make all levels of exclusive data sets and perform sample equalization and expansion operations. After many times of multi-level retraining, the accuracy of text recognition reached 96%, and the form text recognition operation was completed.

Text recognition result：报告编号： Report No.

Figure 4. Text recognition after retraining Tesseract

G. Electronic Form Reconstruction

At last, perform the reconstruction operation of the electronic version of the EXCEL file of the form, use the structure information of the form based on XML analysis and the reconstruction number of the cell image to complete the reconstruction of the form electronic file and fill in the identification content of the corresponding cell. Finally, the recognition of the paper version of the form file and the reconstruction of the electronic form file are completed.

III. EXPERIMENTS

The experiment designed in this paper verifies the research content of this paper. The first is the performance comparison experiment between the lightweight U-Net network model and the original U-Net network model in cell image detection and positioning. First, create a network training and test data set, analyze the main cell shape types, and create a sample balanced data set according to the proportion of the number. The data set has a total of about 2000 cell images. The test data set accounts for about 0.2, and the proportion of the training data set is about 0.8. The experimental results are as follows.

TABLE I. U-NET LIGHTWEIGHT VERIFICATION EXPERIMENT

Models	Performance Comparison		
	Parameters	***Calculation Amount***	***Accuracy***
U-Net	879169	About 809 million	96.4%
Lightweight U-Net	84905	About 129 million	94.3%

From the experimental results, it can be found that the performance of the lightweight U-Net network model based on the deep separable convolution technology has been improved. Compared with the original network model, the parameter amount is reduced by about 90%, and the calculation amount is reduced by about 84%. The prediction accuracy rate of the lightweight network model only dropped by 2.1%. Therefore, it is concluded that, compared with the original U-Net network structure model, the lightweight U-Net network model using deep separable convolution technology will sacrifice certain network model prediction accuracy, but will greatly reduce the amount of parameters and the amount of calculation increases the computational efficiency of the network model. Therefore, the lightweight technology of the network model based on deep separable convolution has important practical production value.

The second experiment is to verify whether the Tesseract neural network can meet the accuracy requirements for Chinese, English, numbers, symbols, and mixed types of text information after retraining. Due to the complexity of recognizing text type information, this article uses a multi-level retraining method to gradually retrain Tesseract to continuously improve its ability to recognize text information. Tesseract was gradually retrained for many times in the order of English recognition, English number mixed recognition, English number symbol mixed recognition, Chinese recognition, and multiple types of random mixed recognition. Based on 3000 commonly used Chinese vocabulary, commonly used English vocabulary, numbers and symbols, a test dataset is produced, and the text recognition accuracy of the test on this dataset reaches 96%.

The third experiment is to verify the feasibility of the table recognition algorithm proposed in this paper. First, we collected 100 documents containing different types of forms and corresponding XML files. The content of the forms contained texts that were randomly mixed in Chinese, English, numbers, and symbols. This experiment has two evaluation criteria, the first is whether the spreadsheet can be reconstructed, and the second is whether the cell content recognition is correct. Among them, the successful reconstruction of the spreadsheet means that the table area is detected and extracted successfully, and the table structure information is analyzed correctly and the coordinate system conversion is correct. The correct recognition of cell content

means that the text detection based on the lightweight U-Net network is correct, and the Tesseract text recognition accuracy after retraining meets the requirements.

TABLE II. ALGORITHM FEASIBILITY VERIFICATION EXPERIMENT

	Evaluation Standard	
	Reconstructions	***Accuracy of Cell Text Recognition***
Accuracy	100/100	94.1%

It can be seen from the experimental results that the XML-based table structure analysis method proposed in this paper can accurately analyze the table structure information with a resolution of 100/100. At the same time, the multiple and multi-level retraining methods proposed in this paper can enable Tesseract to gradually improve the recognition ability of text information, and finally can successfully recognize complex types of text information with a recognition accuracy of 94.1%.

IV. CONCLUSION

The table area detection algorithm in this paper cleverly uses the feature difference between the table and other information in the image to specify the table outline filtering conditions, and successfully realizes the detection and extraction of the table area. This algorithm has the advantages of simple implementation and faster detection speed. The form structure analysis algorithm proposed in this paper makes full use of the prior information of the form file, that is, the XML file is used to accurately analyze the structure information of the form and the coordinate information of the cell, and the corresponding cell image is successfully extracted after the coordinate system conversion. At the same time, in order to reduce the influence of frame and other factors on the accuracy of text recognition, this paper uses the lightweight U-Net model to achieve pixel-level segmentation of the text area in the cell image, and the text area detection accuracy is 94.2%. At the same time, using deep separable convolution to lighten the U-Net model, the parameter amount and calculation amount are reduced by more than 80%, and the performance of the network model is greatly improved on the basis of ensuring the detection accuracy. In the text recognition algorithm, this article uses multiple and multi-level retraining methods to successfully enable Tesseract to have the ability to recognize complex types of text with an accuracy of 96%. In summary, the research in this article is conducive to the batch digital processing of paper-based form files, can greatly reduce the amount of manual labor, reduce the error rate of file processing, and can greatly promote the development of the digital process of enterprises.

ACKNOWLEDGMENT

Thank you for the work of the organizer and related staff of this academic conference. At the same time, I would like to thank my teachers and classmates for their support to my research work.

This work was funded by National Key R&D Program of China (No.2017YFF0210501) and Major scientific and technological innovation platform construction in Xi'an (No. 201809163CX4JC5)

REFERENCES

[1] Zheng Yefeng, Liu Changsong, Ding Xiaoqing, et al. Table frame based on directed singly connected chain Detection Algorithm[J].Journal of Software,2002(4):790-796.

[2] GILANI A, QASIM S R,MALIK I,et al. Table Detection Using Deep Learning[C]// 2014 14th IAPR International Conference on Document Analysis and Recognition.Kyoto: IEEE, 2017:771-776.

[3] SUN N,ZHU Y,HU X.Faster R-CNN Based Table Detection Combining Corner Locating[C]//2019 International Conference on Document Analysis and Recognition (ICDAR).Sydney:IEEE, 2019:1314-1319.

[4] REN S,HE K,GIRSHICK R,et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(6):1137-1149.

[5] HUANG Y,YAN Q,LI Y, et al.A YOLO-Based Table Detection Method[C]//2019 International Conference on Document Analysis and Recognition (ICDAR)

[6] REDMON J,FARHADI A.YOLOv3: An Incremental Improvement[J]. arXiv e-prints,2018:1804.02767.

[7] SIDDIQUI S A,KHAN P I, DENGEL A,et al.Rethinking Semantic Segmentation for Table Structure Recognition in Documents[C]//2019 International Conference on Document Analysis and Recognition (ICDAR).Sydney :IEEE 2019: 1397-1402.

[8] SIDDIQUI S A,FATEH I A,RIZVI S T R,et al.DeepTabStR: Deep Learning Based Table Structure Recognition[C]// 2019 International Conference on Document Analysis and Recognition (ICDAR)., Sydney :IEEE, 2020: 1403-1409.

[9] QASIM S R,MAHMOOD H,SHAFAIT F.Rethinking Table Recognition using Graph Neural Networks[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney :IEEE, 2019:142-147.

[10] PRASAD D,GADPAL A,KAPADNI K,et al.CascadeTabNet: An Approach for End to End Table Detection and Structure Recognition From Image-Based Documents[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops Seattle:IEEE, 2020:2439-2447.

[11] CAI Z,VASCONCELOS N.Cascade R-CNN: Delving Into High Quality Object Detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City :IEEE, 2018:6154-6162.

[12] SUN K,XIAO B,LIU D,et al.Deep High-Resolution Representation Learning for Human Pose Estimation[J]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach:IEEE, 2019:5686-5696.

[13] SIMONYAN K,ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[14] ANON. IEEE Conference on Computer Vision And Pattern Recognition(cvpr2020)[J].JournalofIntelligent Systems,2019,14(6):1137.

[15] HUANG G,LIU Z,LAURENS V D M,et al.Densely Connected Convolutional Networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu :IEEE,2017:4700-4707.

[16] GRAVES A.Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks [C]// International Conference on Machine Learning.

[17] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 3104−3112.

[18] SHI B,BAI X,YAO C.An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2016,39(11): 2298-2304.

[19] Wu Yi. XML practical technology self-study classic [M]. Beijing: Tsinghua University Press, 2015.

[20] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Munich, Germany: Springer, 2015: 234-241.

[21] Sifre L，Mallat S.Rigid - Motion Scattering for Texture Classification［EB/OL］.［2021 - 02 - 25］.https://arxiv.org/ pdf/1403.1687.pdf.

[22] SMITH R.An Overview of the Tesseract OCR Engine[C]// Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on Document Analysis and Recognition(ICDAR 2007). Parana:IEEE, 2007(2): 629-633.