JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, 1305-1315 (2022)
DOI: 10.6688/JISE.202211_38(6).0011

Deformable Convolutional Neuron Network Model for Detecting Tables and Columns from Document Images*

WEN-TIN LEE+ AND CHUAN-CHUN HUANG
Department of Software Engineering and Management
National Kaohsiung Normal University
Kaohsiung, 802 Taiwan
E-mail: wtlee@mail.nknu.edu.tw; 610877105@o365.nknu.edu.tw

Tables are usually used to present the main points of a document so that readers can quickly understand the content of the document. This study proposed a novel deformable convolutional neural network model for table detection to identify and extract tables from electronic document images. The model can perform table detection and table structure recognition at the same time, and more effectively detect the location of tables and columns. The proposed model is evaluated using Marmot extended dataset and the experimental results show that the table detection cycle is reduced, the computation time is shortened, and the overall efficiency is improved. Compared with other studies, the proposed model has achieved better table detection results in terms of precision, recall, and F1-score.

Keywords: table detection, table structure recognition, column identification, deformable CNN, deep learning

1. INTRODUCTION

In recent years, there has been a dramatic increase in the number of document images uploaded by users using cameras or scanners. In such cases, how to quickly obtain the information and find out the key content from document images is an important topic [1]. The tables in documents usually present important information in a simple, standard, and structured way. These tables in document images are usually processed manually to find the required information, resulting in a waste of labor cost and processing time. Table detection and recognition is an important step in analyzing and understanding the content of the entire document image. Although some table identification methods have been proposed to speed up table detection and table structures recognition, it is still a challenging problem due to various layouts and visual appearance of different tables.

Table detection may also encounter the interference problem of recognizing graphs and pictures with small differences [2]. Solving the above problems is an important and challenging work, so the main purpose of this work is to propose a new table detection framework to improve performance, precision, and recall rate on multiple well-known public data sets.

Before the advent of deep learning and convolutional neural network, most of the table detection model was based on heuristics and metadata training. A convolutional neural network (CNN) is a feed-forward neural network, which is composed of one or more convolution layers, pooling layers, and fully connected layers. The convolution layer, pooling layer, and fully connected layer are used to extract feature value in images and classify

Received November 1, 2021; revised January 7, 2022; accepted February 7, 2022.
Communicated by Shin-Jie Lee.
* This research was sponsored by the Ministry of Science and Technology in Taiwan under Grands No. MOST 110-2221-E-017-001.

the features. Yann [3] propose the LeNet-5 model which is a gradient-based learning CNN structure called Graph Transformer Networks, which is used to train multi-module systems to minimize an overall performance measure. Because CNN has excellent performance on the spatial structure, it has become the best way to train a stabilization model and is widely used in the fields of textual analysis, image recognition, and natural language processing (NLP).

The receptive fields of all neurons in a given convolutional layer are the same which is problematic for layers located on top of the hierarchy where different objects may appear at arbitrary scales along with arbitrary transformations. The deformable convolutional neural network (Deformable CNN, DCNN) has derived from CNN to break the neuronal limitation. Deformable convolutional layers' neurons were not limited to a predefined receptive field [4]. Each neuron can alter its receptive fields by adjusting the eigenvalues according to its input via generating explicit offsets. This allows the convolutional layers to adapt to different scales and transformations by adjusting their receptive fields. Since tables can be presented at random scales with any transformation (orientation, position, size, *etc.*), the deformable convolution operations are particularly useful for table detection tasks.

In this study, we propose TD-DCNN a deformable convolutional neural network model for detecting tables in document images. TD-DCNN can automatically extract table features from document images, which eliminates the limitation of the general applicability of heuristic learning methods [5]. By replacing the traditional convolutional neural network with a deformable convolutional neural network (D-CNN), TD-DCNN identifies the location and size of tables from document images with more accurate results. In terms of performance improvement, this work designs the task flow by finding the operational relationship between table detection and column recognition to simplify the execution of the entire system.

2. RELATED WORK

Work in a number of fields has made its mark on our research. Our approach has drawn upon several ideas from data-driven models and deep learning techniques.

Silva *et al.* [6] proposed a data-driven model based on hidden Markov models (HMMs) to improve table recognition accuracy. Their approach establishes joint probability functions by using a sequential interpretation of the visual page elements with the rows in the hidden layer to locate tables. Kasar *et al.* [7] proposed a method to detect table regions in document images by identifying line separators and properties of columns and rows. Although it does not use heuristic rules or pre-defined parameters, it needs to rely on clear table borders. They used an SVM classifier to verify features by grouping intersecting horizontal and vertical lines. Tran *et al.* [8] proposed a method based on regions of interest and the spatial arrangement of extracted text blocks. The proposed method is applied to the ICDAR 2013 data set [9] to detect tables from document images.

There are many limitations in those table detection methods, including the inability to identify incomplete and compact tables by detecting horizontal or vertical lines. During the feature identification period, its performance cannot be effectively improved because it relies heavily on labor costs.

Hao *et al.* proposed the first table recognition framework based on deep learning [10], and its input is limited to PDF files. This framework uses heuristic rules that define features and the unit data of the PDF documents for training. Some predefined set of rules are used

to calculate region proposals passed to CNN to check whether the region proposals belong to the table region of the document.

DeepDeSRT [11] uses deep learning for table detection and table structure recognition to identify rows, columns, and cell positions in the detected tables. By improving the pre-training model of Faster R-CNN [12], and enhancing the FCN semantic segmentation model pre-trained on Pascal VOC 2011 [13, 14], the transfer learning function is performed to find out the position of each table rows, columns, and sub-tables.

Kavasidis *et al.* [5] propose a system that generates saliency maps using Fully-Convolutional Networks (FCN) [13] and use Conditional-Random Fields on top of it to apply additional constraints on the generated masks. They also train C binary classifiers (where C is the number of classes) to supplement the performance of the saliency network. They used dilated convolutions proposed by Yu and Koltun [15] to increase the effective receptive field of the layers without any loss of resolution. The method was trained using a semantic database of 50442 images created by a web crawler.

Siddiqui *et al.* [16] proposed a deep deformable CNN model for table detection. It uses deviation values to adapt its receptive field according to its input and expands coverage area when encountering tabular data to improve the accuracy and performance of table detection. They evaluated the model on the publicly available datasets of ICDAR-2013, ICDAR-2017 POD, UNLV, and Mormont to show that their approach can effectively improve the operating efficiency of the model.

Paliwal *et al.* [17] demonstrated a end-to-end deep learning model – TableNet which was based on pre-trained VGG-19, to detect tabular regions and column masks. They combine the two tasks into one model by using similar content in the first half of the two models and adding dropout to strengthen the stability of the eigenvalues of different studies. The proposed method has been verified on the ICDAR-2013 data set, showing that it can effectively improve the operating efficiency and optimize the operating volume of the model. Furthermore, ROI-pooling is a core component for all region-based detection methods [12, 18]. The pooling layer converts feature maps of arbitrary size to a fixed volume to be fed to the final classification. However, it has fixed-receptive field problems similar to the one in a conventional CNN. Table 1 compares our model with the above-mentioned work, except DeCNT, other models use traditional convolutional neuron network for detecting tables. DeCNT, however, cannot detect table columns or more complex structures due to the restriction caused by the type of input.

Table 1. Comparison of work on table detection.

Model	CNN type	Dropout
Our model	Deformable	Two
DeepDeSRT [11]	Traditional	None
DeCNT [16]	Deformable	None
TableNet [17]	Traditional	Two

3. PROPOSED METHODOLOGIES

Table detection using deep learning and convolutional neural networks requires complex steps to train from the beginning This study simplifies the steps to propose a novel

table detection and column identification model. The deformable convolutional neural network can provide better feature value extraction and detect tables of arbitrary layouts that exist in various place in the document. Therefore, we use deformable convolutions to design the table detection model named TD-DCNN. Fig. 1 shows the structural model of TD-DCNN which uses Deformable CNN to extract the table features in the input file. To avoid the problem of overfitting, dropout is used to normalize the feature value to prevent the extracted features from affecting each other. Table Detection uses fractionally stridden convolutions to yield the mask for table regions. The branch of table column identification is performed at the same time as the table detection which performs another convolution and dropout to extract column structure features. It adds a deformable convolution layer to extract column structural features inside the table. Structure Recognition uses fractionally stridden convolutions to yield the mask for column regions. Finally, the recognition results are generated using deconvolution.

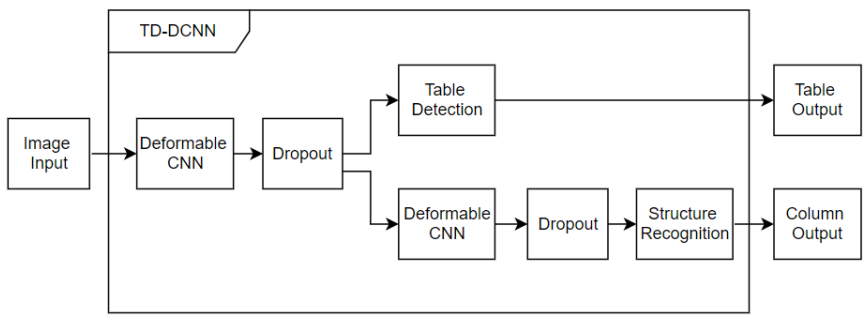


Fig. 1. Structural model of TD-DCNN.

Fig. 2 shows the details of each stage. In stage (a), we change traditional convolutional layers and traditional pooling layers in the VGG-19 neural network model to a deformable structure. The 3×3 feature extraction matrix is used to set the offset-channel and generate 9 feature values. The feature values are added to the original convolution parameters to obtain the new convolution layer parameters. The pooling layer also uses the feature map according to the originally RoI-pooling ratio to obtain new parameters.

In Stage (b), the fully connected layers of VGG-19 are replaced with an 8×8 convolution layer (Conv6). The convolution layer uses a dropout layer with dropout = 0.8 to prevent the model from overfitting. After that, the model will be divided into two branches: the table detection and the column identification.

Stage (c) performs table detection. Before performing deconvolution to output the picture, a 1×1 convolution layer named Conv7_table_conv is used to extract the feature to optimize table results. The output of the Conv7_table_conv convolution layer is also up-scaled using fractionally stridden convolutions and is appended with the Conv4 pooling layer. Similarly, the Conv3 pooling is appended to the up-scaled output from the previous layer. Finally, the final feature map is upscaled to meet the original image dimensions, and the mask and score of the table are generated as the table detection results.

Stage (d) performs columns identification. There is an additional convolution layer named Conv7_column_conv with a deformable convolution function and a dropout layer

with the ame dropout probability to extract the feature in the deeper layer. The feature maps are up-scaled using fractionally stridden convolutions after an 8×8 convolution (Conv8_column_conv) layer. The up-scaled feature maps are combined with the Conv4 pooling layer and the combined feature map is up-scaled and combined with the Conv3 pooling layer of the same dimension. After this layer, the feature map is up-scaled to the original image and then complete the mask and score of the column detection.

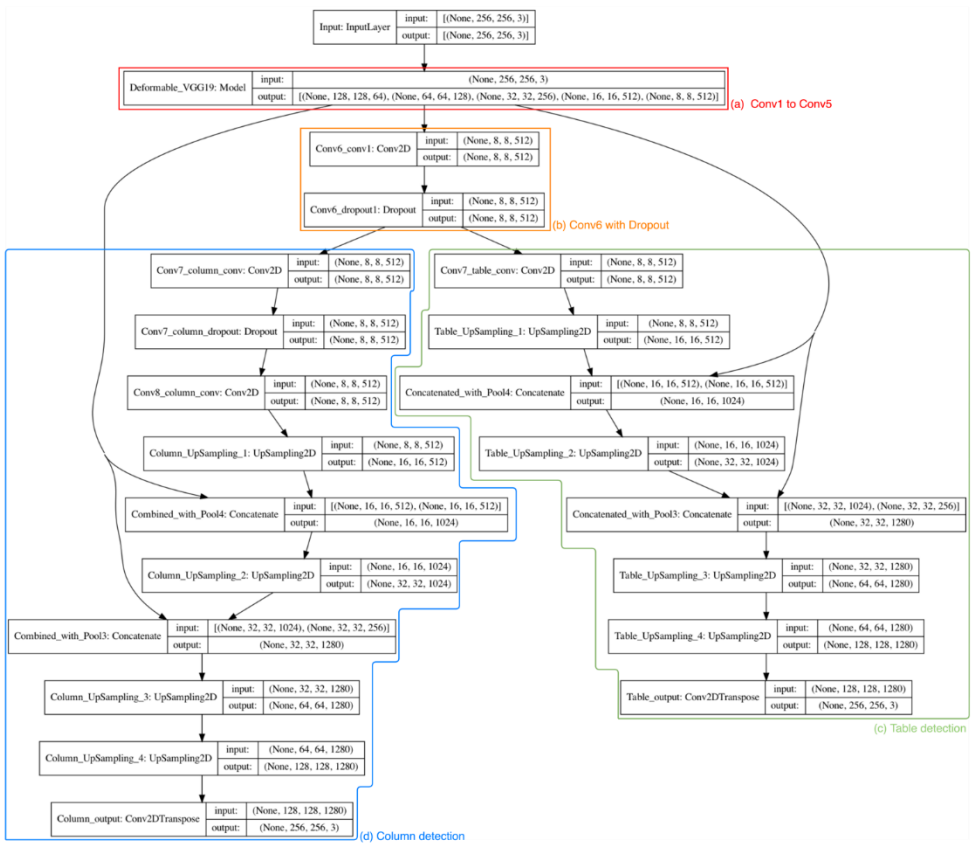


Fig. 2. Detail stages of TD-DCNN.

Deformable convolution can shorten the operating time of the convolutional layer and increase the range of feature value extraction. The input of this model is an image, so only documents with tables need to be scanned into image files to meet the demand of general utilization. Since word detection and extraction will produce some noise, another dropout was added to randomly discard certain types of data to reduce the amount of noise. However, this may cause the loss of important feature values. By supplementing the features of the upper convolutional layer, the feature values can be retained as much as possible.

The data set used for table recognition in the current research field is the ICDAR competition data set. To train the TD-DCNN model, we used the Marmot extended dataset which was collated and constructed from the ICDAR 2013 dataset by Paliwal *et al.* [18].

The Marmot dataset consists of two parts, Chinese and English. The Chinese texts are from more than 120 e-books in different disciplines, and the English texts are from the scientific and technical academic papers on the Citeseer website. There are 1016 documents containing tables, including Chinese and English documents. Most Chinese documents have a single-column layout, while English documents have a mixture of single-column and double-column layouts. In addition to image files, Marmot also stores a tree structure of all document content layouts, with internal nodes including lines, paragraphs, tables, *etc.* in the text.

To provide basic semantic information to the model, tesseract OCR is used to color-code the text of the article image. Each type will use a different color as an annotation and color the boundary of a word to convey semantic and spatial information to the neural network. Shadows will be added to two similar categories to enhance the discrimination and facilitate the elimination of false detections.

4. EXPERIMENTAL RESULTS

This section describes the experiments evaluated on the Marmot dataset [18] for table and column detection. Python, TensorFlow, and Anaconda are used to develop, manage and deploy programs of TD-DCNN. The hardware used for the experiment is a 2.3GHz, 8-core Intel Core i9 processor, 16 GB 2667 MHz DDR4 memory, and AMD Radeon Pro 5500M 4 GB graphics card.

Table 2. Results for table detection.

Model	Recall	Precision	F1-score	Time	Epochs
TD-DCNN	**0.973**	0.967	**0.969**	**477 sec**	**1500**
TableNet [17]	0.969	0.967	0.968	491 sec	3000
DeepDeSRT [11]	0.962	**0.974**	0.967	N/A[*1]	30000
DeCNT[16]	0.946	0.849	0.895	N/A[*2]	N/A[*2]

[*1] DeepDeSRT does not mention the time required for the operation. (N/A means Not Applicable)
[*2] DeCNT does not mention the cycle and time required for the operation.

The model performance is evaluated based on the recall, precision, and F1-score. The experimental results are computed and compiled in Table 2. Compared with TableNet, DeepDeSRT, and Tran *et al.*, TD-DCNN has better results with recall = 0.973, precision = 0.967, and F1-score = 0.969. Furthermore, TD-DCNN only needs 1500 iterations to surpass the results of 30,000 iterations of DeepDeSRT and 3000 iterations of TableNet, indicating that the process proposed in this study has improved operating performance. Because DeepDeSRT was built with two different models, it needs to run another 3000 iterations. TableNet and TD-DCNN are both built-in integrated design models, so there is no need to train again.

Table 3 shows the table columns identification results. TableNet can handle the scope issues, but the feature noise increases and the precision decreases. DeepDeSRT has a good precision value, but its recall rate is lower than other methods due to the limitation of detection scope. The design of DeCNT is mainly to change the convolution model structure in object identification to a deformable structure, but due to the lack of available data

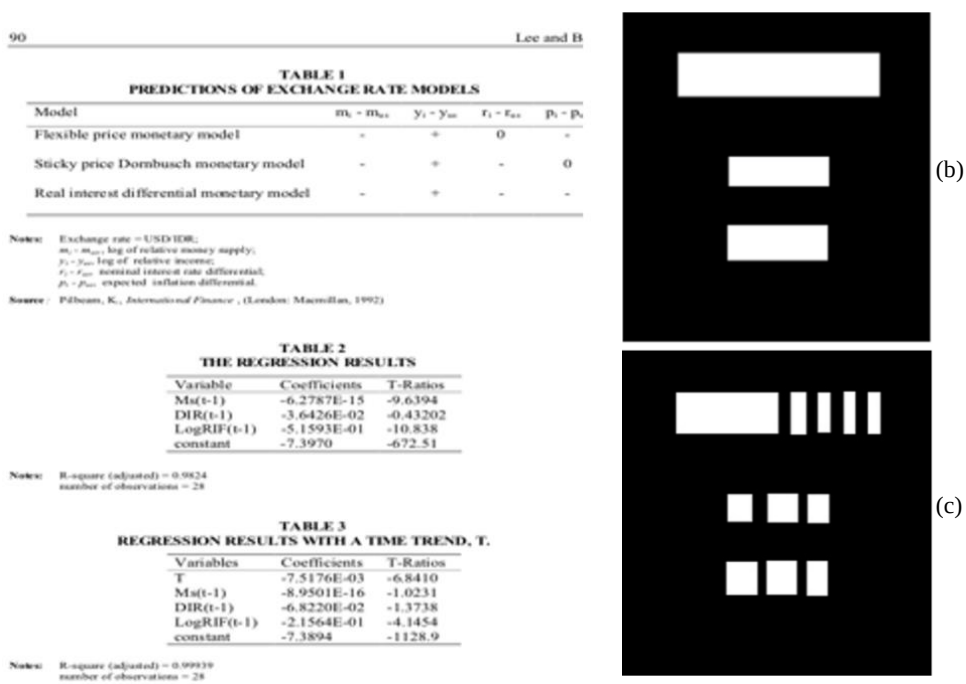
sets at that time, it cannot effectively identify the complex and variable structures and columns in the table. Finally, TD-DCNN uses deformable convolution to increase the range as much as possible to maintain the cleanliness of feature values and improve the performance of the system.

Table 3. Results for table columns identification.

Model	Recall	Precision	F1-score	Epochs
TD-DCNN	**0.9343**	0.9095	**0.922**	1500
TableNet	0.9250	0.9055	0.916	3000
DeepDeSRT	0.8736	**0.9593**	0.914	3000
DeCNT	N/A*			

* DeCNT cannot perform table structure recognition.

Fig. 3 shows an image of the Marmot extended data set. Fig. 3 (a) is a raw document image with three tables, Figs. 3 (b) and (c) are the table mask and the column mask of Fig. 3 (a), respectively. Fig. 4 is the output generated by TableNet, and Fig. 5 is the output generated by TD-DCNN. From the output results of the two research experiments, Figs. 5 (a) and (b) show that the content near the boundary in the table is noisy due to the detection bias of deformable convolution. In column identification, category segmentation has many inaccurate problems. In TableNet, Fig. 4 (b) shows a blocky picture, which makes the values of false-positive rise sharply and the precision rate becomes worse. Fig. 5 (b) shows that despite the noise, TD-DCNN has generated a separate block structure.



90 Lee and B

TABLE 1
PREDICTIONS OF EXCHANGE RATE MODELS

Model	$m_t - m_{us}$	$y_t - y_{us}$	$r_t - r_{us}$	$p_t - p_t$
Flexible price monetary model	-	+	0	-
Sticky price Dornbusch monetary model	-	+	-	0
Real interest differential monetary model	-	+	-	-

Notes: Exchange rate = USD/IDR;
$m_t - m_{us}$, log of relative money supply;
$y_t - y_{us}$, log of relative income;
$r_t - r_{us}$, nominal interest rate differential;
$p_t - p_{us}$, expected inflation differential.

Source: Pilbeam, K., *International Finance*, (London: Macmillan, 1992)

TABLE 2
THE REGRESSION RESULTS

Variable	Coefficients	T-Ratios
Ms(t-1)	-6.2787E-15	-9.6394
DIR(t-1)	-3.6426E-02	-0.43202
LogRIF(t-1)	-5.1593E-01	-10.838
constant	-7.3970	-672.51

Notes: R-square (adjusted) = 0.9824
number of observations = 28

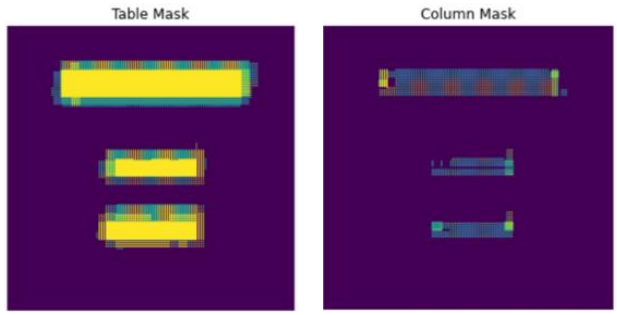
TABLE 3
REGRESSION RESULTS WITH A TIME TREND, T.

Variables	Coefficients	T-Ratios
T	-7.5176E-03	-6.8410
Ms(t-1)	-8.9501E-16	-1.0231
DIR(t-1)	-6.8220E-02	-1.3738
LogRIF(t-1)	-2.1564E-01	-4.1454
constant	-7.3894	-1128.9

Notes: R-square (adjusted) = 0.99939
number of observations = 28

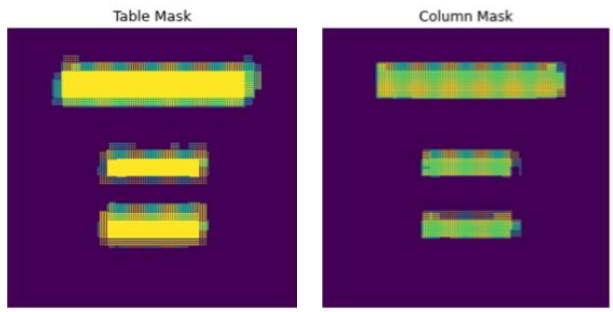
(a) (b) (c)

Fig. 3. The image of Marmot extended data; (a) Raw document image; (b) Table mask; (c) Column mask.



(a) (b)

Fig. 4. TableNet output; (a) Table mask; (b) Column mask.



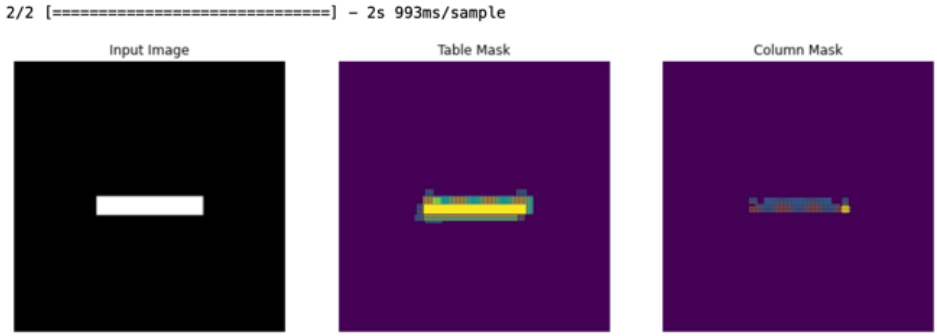
(a) (b)

Fig. 5. TD-DCNN output; (a) Table mask; (b) Column mask.

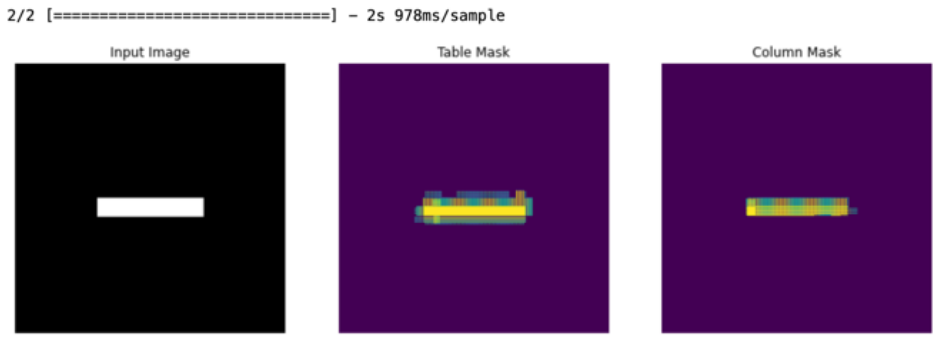
In terms of performance optimization, as the number of training iterations reduces, the execution time is improved. Compare this study with TableNet, the image processing speed of TableNet is 0.993 sec/image, while the TD-DCNN model can be increased to 0.978 sec/image due to the deformable convolution method. As shown in Fig. 6, the performance of the TD-DCNN model is better than TableNet.

This study uses deformable convolutional neural networks to transform the feature value extraction into conditional changes. When processing the table, the recognition area will be enlarged to shorten the execution cycle and time. At the same time, when it is not a table, the recognition range is narrowed to perform detailed feature extraction to reduce the chance of missing features and improve accuracy. Finally, the previous feature values are added to the reverse convolution to avoid distortion caused by excessive expansion.

In terms of performance evaluation, the TD-DCNN proposed in this study proposes an integrated model that can perform two tasks concurrently by analyzing whether there is a correlation between table position and column structure. Compared with the DeepDeSRT design that performs two tasks separately, the number of iterations and execution time can be effectively reduced. Compared with the feature extraction limitation caused by the use of traditional convolution in TableNet, which affects the detection accuracy and requires a lot of execution time, the deformable convolutional neural network used by TD-DCNN reduce the number of extraction times by enlarging the recognition area when encountering a table, reducing the execution time, and narrowing down non-tabular areas to provide better feature extraction, which can further improve table and column detection accuracy.



(a) TableNet



(b) TD-DCNN

Fig. 6. Running time comparison.

5. CONCLUSIONS

This study proposes a deep learning model named TD-DCNN to identify the table position and table column of the image files. Existing approaches take table detection and structure recognition as two separate problems to be solved independently. TD-DCNN is a model that handles these two tasks at the same time by utilizing the inherent interdependence between table detection and table column identification. TD-DCNN adds deformable convolution to increase the table detection range and keep the feature stable during the detection process.

This study compared TD-DCNN with the methods proposed by TableNet, DeepDeSRT, DeCNT, and Tran *et al.* The experimental results show that TD-DCNN uses deformable convolution to improve the accuracy of table detection and table column identification, effectively reduce the calculation cycle required for detection, and improve the overall efficiency.

Although TD-DCNN has superior results compared with several modern research models, this work only uses the Marmot extension dataset for evaluation. Tables come in many forms and we cannot guarantee that all table detections for documents will work correctly. In the future, we plan to propose a third branch to train TD-DCNN to recognize the information in the table. To address the research limitation, it may be useful to use more complex tables and forms, such as merged cells, grids, background colors, document clarity, *etc.* to obtain better model performance.

REFERENCES

1. B. Coüasnon and A. Lemaitre, "Recognition of tables and forms," *Handbook of Document Image Processing and Recognition*, D. Doermann and K. Tombre, eds., Springer, London, 2014, pp. 647-677.
2. D. W. Embley, M. Hurst, D. Lopresti, and G. Nagy, "Table-processing paradigms: a research survey," *International Journal of Document Analysis and Recognition*, Vol. 8, 2006, pp. 66-86.
3. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, Vol. 86, 1998, pp. 2278-2324.
4. J. Dai *et al.*, "Deformable convolutional networks," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 764-773.
5. I. Kavasidis *et al.*, "A saliency-based convolutional neural network for table and chart detection in digitized documents," *arXiv Preprint*, 2018, arXiv:1804.06236.
6. A. C. e Silva, "Learning rich hidden markov models in document analysis: Table location," in *Proceedings of IEEE 10th International Conference on Document Analysis and Recognition*, 2009, pp. 843-847.
7. T. Kasar, P. Barlas, S. Adam, C. Chatelain, and T. Paquet, "Learning to detect tables in scanned document images using line information," in *Proceedings of IEEE 12th International Conference on Document Analysis and Recognition*, 2013, pp. 1185-1189.
8. D. N. Tran, T. A. Tran, A. Oh, S. H. Kim, and I. S. Na, "Table detection from document image using vertical arrangement of text blocks," *International Journal of Contents*, Vol. 11, 2015, pp. 77-85.
9. A. C. e Silva, "Metrics for evaluating performance in document analysis: application to tables," *International Journal on Document Analysis and Recognition*, Vol. 14, 2011, pp. 101-109.
10. L. Hao, L. Gao, X. Yi, and Z. Tang, "A table detection method for pdf documents based on convolutional neural networks," in *Proceedings of IEEE 12th IAPR Workshop on Document Analysis Systems*, 2016, pp. 287-292.
11. S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed, "Deepdesrt: Deep learning for detection and structure recognition of tables in document images," in *Proceedings of IEEE 14th IAPR International Conference on Document Analysis and Recognition*, Vol. 1, 2017, pp. 1162-1167.
12. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, Vol. 28, 2015, pp. 91-99.
13. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.
14. M. Everingham, L. van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, Vol. 88, 2010, pp. 303-338.
15. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv Preprint*, 2015, arXiv:1511.07122.
16. S. A. Siddiqui, M. I. Malik, S. Agne, A. Dengel, and S. Ahmed, "Decnt: Deep defor-

mable cnn for table detection," *IEEE Access*, Vol. 6, 2018, pp. 74151-74161.
17. S. S. Paliwal, D. Vishwanath, R. Rahul, M. Sharma, and L. Vig, "Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images," in *Proceedings of IEEE International Conference on Document Analysis and Recognition*, 2019, pp. 128-133.
18. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, 2016, pp. 1137-1149.

Wen-Tin Lee (李文廷) received his Ph.D. degree in Computer Science and Information Engineering from National Central University, Taiwan, in 2008. Lee is currently an Associate Professor in the Department of Software Engineering and Management at National Kaohsiung Normal University. His research interests include software engineering, service-oriented computing, and deep learning.

Chuan-Chun Huang (黃傳鈞) is currently a master student in the Department of Software Engineering and Management at National Kaohsiung Normal University, Taiwan. His research interests include deep learning, artificial intelligence, and image processing.