2019 International Conference on Document Analysis and Recognition (ICDAR)

DoT-Net: Document Layout Classification Using Texture-based CNN

Sai Chandra Kosaraju
Department of Computer Science
Kennesaw State University
Marietta,USA
skosara1@students.kennesaw.edu

Mohammed Masum
Analytics and Data Science Institute
Kennesaw State University
Kennesaw,USA
mmasum@students.kennesaw.edu

Nelson Zange Tsaku
Department of Computer Science
Kennesaw State University
Marietta,USA
ntsaku@students.kennesaw.edu

Pritesh Patel
GE power
Atlanta,USA
pritesh.patel1@ge.com

Tanju Bayramoglu
GE power
Atlanta,USA
tanju.bayramoglu@ge.com

Girish Modgil
GE power
Atlanta,USA
girish.modgil@ge.com

Mingon Kang*
Department of Computer Science
Kennesaw State University
Marietta,USA
mkang9@kennesaw.edu

***Abstract*—Document Layout Analysis (DLA) is a segmentation process that decomposes a scanned document image into its blocks of interest and classifies them. DLA is essential in a large number of applications, such as Information Retrieval, Machine Translation, Optical Character Recognition (OCR) systems, and structured data extraction from documents. However, identification of document blocks in DLA is challenging due to variations of block locations, inter- and intra- class variability, and background noises. In this paper, we propose a novel texture-based convolutional neural network for document layout analysis, called DoT-Net. DoT-Net is a multiclass classifier that can effectively identify document component blocks such as text, image, table, mathematical expression, and line-diagram, whereas most related methods have focused on the text vs. non-text block classification problem. DoT-Net can capture textural variations among the multiclass regions of documents. Our proposed method DoT-Net achieved promising results outperforming state-of-the-art document layout classifiers on accuracy, F1 score, and AUC. The open-source code of DoT-Net is available at https://github.com/datax-lab/DoTNet.**

***Index Terms*—document layout analysis, dilated CNN, texture-based document analysis**

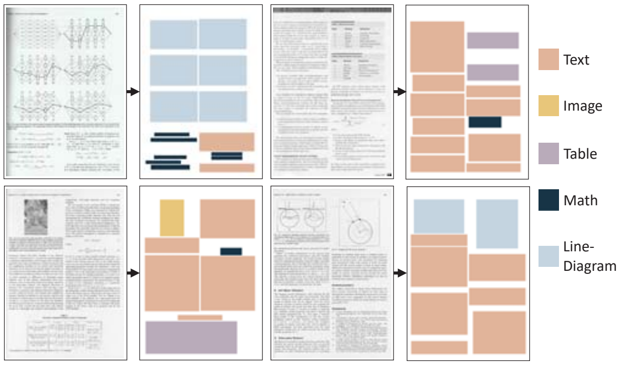


Fig. 1: **Texture-based CNN for document layout classification (DoT-Net)**. We present a novel texture-based CNN to classify document blocks such as text, image, table, math, and line-diagram for document layout analysis.

I. Introduction

Document Layout Analysis (DLA) is a segmentation process that decomposes a scanned document image into its blocks of interest and classifies them, e.g. text, image, table, mathematical expression, and line-diagram [1]. DLA leads to a large number of applications, such as information retrieval, machine translation, Optical Character Recognition (OCR) systems [2], and structured data extraction from documents [3]. However, classification of document blocks in DLA is challenging, due to variation of block locations, inter- and intra- class variability, and background noise.

DLA mainly consists of three procedures: (1) detecting document blocks of interest, (2) extracting features, and (3) classifying the blocks. Traditional block detection methods of top-down, bottom-up, and hybrid approaches have been used to localize document blocks [4]. Then, features are extracted from the blocks by using block-based, pixel-based, or connected component-based techniques [1]. In particular, pixel-based features include entropy, gradient shapes, and contrasts, whereas texture-based features contain size, shape, stroke width, and positions of the blocks. The extracted features are then introduced into a machine learning algorithm for classifying the document blocks. In this study, we focus on classification of document blocks where localized document blocks are given.

A number of machine learning algorithms have been applied for document layout classification with features that describe characteristics of document blocks. Gradient shape features were generated to represent textual patterns and introduced to a Support Vector Machine (SVM) classifier for classifying text blocks from non-text blocks [5]. A Multilayer Perceptron (MLP) was trained with Histogram of Oriented Gradients (HOG) descriptor to classify text and non-text document blocks [6]. An adaptive boosting (Adaboost) decision tree was applied for classification between text and non-text regions

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00168

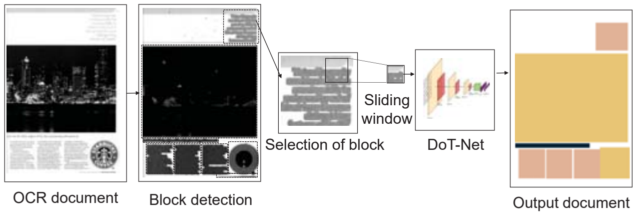


Fig. 2: Overview of proposed algorithm. DoT-Net classifies tile images of a block and the final classification decision of the block is made by majority vote.

using features extracted by the connected component approach [7].

Recently, deep learning has been widely explored in document layout classification. A feed forward neural network was trained with textural and statistical features extracted by processing a mask function across document images for the text vs. non-text classification [8]. A fast Convolutional Neutral Network (CNN) based document layout analysis was introduced, where two one-dimensional projection of images were considered to train the model [9]. To identify complex document layouts, a CNN architecture that learns a hierarchy of features from a raw image was proposed for the document image classification [10]. A Deep CNN architecture was applied for classification, where CNNs were extensively used for both feature extraction and model training process [11].

Most DLA studies have been mainly focused on a binary classification between text and non-text blocks. Meanwhile, non-text types of blocks such as table, image, mathematical expression, and line-diagram also play an important role in applications of DLA. However, there has been a little published research for classifying specific non-text types of blocks. For instance, a gradient boosted decision tree based classification model was adopted to recognize layout tables and extract encoded knowledge from the tables [12]. Therefore, multiclass classification approaches may increase the efficiency and the scope of the document layout analysis.

In this paper, we propose a document texture-based CNN (DoT-Net), which can effectively and simultaneously classify multiple classes of document blocks (see Fig. 1). Our main contributions are: (1) adopting a dilated convolutional layer replacing all convolutional layers for the texture based analysis, (2) automatic feature extraction via a deep learning model rather than using explicitly predefined features, and (3) extending to multiclass classification whereas previous methods have typically focused on binary classification of text vs. non-text.

The overall procedures of document layout analysis are as follows. Given a document that has gone through OCR, document blocks are first localized. Then, tile images are generated by sliding a sub-window across a block. Each tile image is classified by DoT-Net. Finally, the document block is classified by majority voting. The procedure is illustrated in Fig. 2.

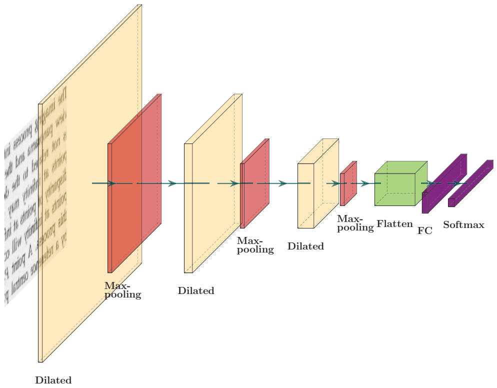


Fig. 3: The architecture of the proposed DoT-Net

II. Methods

In this section, we elaborate our novel texture-based convolutional neural network (DoT-Net). DoT-Net enhances a deep learning architecture for document layout classification. DoT-Net adopts dilated convolutional layers to extract texture patterns from document blocks, which tackle the drawbacks of conventional CNN.

DoT-Net consists of an input layer, three dilated convolutional layers, a flatten layer, a fully connected layer, and an output layer, as shown in Fig. 3. DoT-Net directly takes a sub-window (a tile image of $p \times p$ pixels) of a block rather than predefined features extracted from a block. A tile image is introduced to the input layer, and DoT-Net outputs a class label of document block types. Then, the majority class of tiles is assigned to the document block.

Dilated convolutional layers have been widely used for object localization, as an alternative of conventional convolutional layers [13], [14]. Dilated convolutional layers enlarge field-of-view (texture) of filters without loss of spatial information [15]. In dilated convolutional layers, the numbers of parameters do not increase while enlarging a kernel size, which makes model training computationally efficient. Moreover, dilated convolutional layers trade off context assimilation against computational time [3].

Dilated convolutional layers can capture texture patterns from an image. DoT-Net fully takes advantage of dilated convolutional layers for texture-based analysis in document layout classification. In DoT-Net, conventional convolutional layers are replaced by dilated convolutional layers, where each dilated convolutional layer is followed by a max pooling layer to control layer sizes in between two dilated layers. Without the max pooling layer, the size of the following layer would increase due to enlarged kernels of dilated convolution. Thus, most studies have used dilated convolutional layers as a deconvolutional layer or the last layer following convolutional layers to localize objects in images [16]–[21]. To the best of our knowledge, no studies have adopted dilated convolutional layers while replacing all convolutional layers. Note that most related works in DLA take the input of features extracted from blocks, whereas DoT-Net classifies tile images directly by automatically learning distinguishable texture patterns of document blocks.

III. Experimental Results

We conducted extensive experiments to assess the proposed method, DoT-Net. We used ICDAR document layout analysis dataset [22] that consists of more than 400 annotated documents[1]. The dataset includes fourteen annotated blocks such as text, table, and images. Among them, we considered the five major block categories of text ($n = 1,432$), image (248), table (119), math (91), and line-diagram (82).

We evaluated the performance of our multiclass classifier with existing cutting-edge methods with the following two experiment settings: (1) one-vs.-rest and (2) multiclass classification using either tile or block images extracted from PDF documents. Since most related studies have focused on a binary classification problem between text and non-text blocks, the multiclass classification problem was converted into one-vs.-rest classification problems for the comparison. For each one-vs.-rest classifier, we randomly selected 80 blocks on a positive class and 20 on each of the other classes, so that totally 80 blocks are from the remaining classes. For the multiclass classification experiments, we randomly selected 80 blocks per class.

Moreover, we considered two different types of input data: (1) blocks and (2) tiles of a block. A block was introduced to benchmark methods in block-wise experiments, whereas a tile of a block was an input in tile-wise experiments. We measured the performance based on the input. Document blocks were given from the labeled dataset. Given 80 document blocks per class, we generated tiles by sliding a small window of 100 × 100 pixels and a stride of 30. Finally, 15,000 tiles of 100 × 100 pixels on average were available for per class.

We compared the performance of DoT-Net with five document layout methods of both binary and multiclass classifiers. The benchmark classifiers were included: Feed Forward Networks (FFN) [8], Fast One Dimensional CNN (F1DCNN) [9], Support Vector Machine with Gradient shape features (GSVM) [5], Multilayer Perceptron with HOG features (HOGMLP) [6], and conventional CNN (CNN) [23]. We used 5-fold cross-validation, where 20% of the training data was used as validation data for optimizing hyper-parameters for each experiment. All experiments were repeated ten times for model reproducibility.

Accuracy, F1 score, and Area Under Curve (AUC) were measured to assess the performance of the methods. We obtained true positive (TP), false positive (FP), true negative (TN), and false negative (FN) on each experiment. Accuracy is denoted by the overall prediction accuracy, which was measured by (TP + TN) / (TP + FP + TN + FN). F1 score was calculated by 2 × (precision × recall) / (precision + recall), where precision = TP / (TP + FP) and recall = TP / (TP + FN). Receiver Operating Characteristic (ROC) curves were traced over various thresholds to examine the trade-off between True Positive Rate (TPR = TP / (TP + FN)) and False Positive Rate (FPR = FP / (FP + TN)). Then, an AUC was computed by the area under the ROC curve.

DoT-Net was implemented by Keras with TensorFlow backend. We set the kernel size to 3×3 and dilation rate to 2. A TanH activation function was used for dilated convolution layers with 50 filters. The max pooling layer of size 2×2 with dropout of 0.1 between the each max pooling layer and dilated convolutional layer was used. The fully connected layer with 50 nodes and the softmax layer with 5 nodes were considered. We also applied minibatch training, where each minibatch size was 32. Two hyperparameters, learning rate and weight decay, were optimized automatically by grid search, to minimize the error in validation data for each experiment.

Experimental settings for the five benchmarks are as follows:

1) Feed Forward Networks (FFN): A non-overlapping mask of size 5×5 across a resized input block (256×256 pixels) generated six statistical features of median, mode, entropy, contrast, energy, and homogeneity [8]. The features were input to the feed forward network that consists of an input layer and a hidden layer with 7 nodes and ReLu activations.

2) Fast One-Dimensional CNN (F1DCNN): Fast one-dimensional CNN was proposed to overcome the computational expense of conventional CNNs [9]. Vertical and horizontal projections of one dimensional array were used as an input. The architecture follows two individual 1DCNN tracks, each of which contains sequence of three one-dimensional convolutional layers with kernel size of 3 ×1. Each convoluational layer followed by max pooling with 2 pixels and 0.1 dropout.

3) Support Vector Machine with Gradient shape features (GSVM): GSVM is a block-based classifier. Gradient shape features extracted from a block were introduced to Support vector machine with RBF kernel [5].

4) Multilayer perceptron classifier with HOG features (HOGMLP): HOGMLP is a block-based classifier, where Histogram of gradient shape features of a block (HOG features) were input to Multilayer preceptron [6].

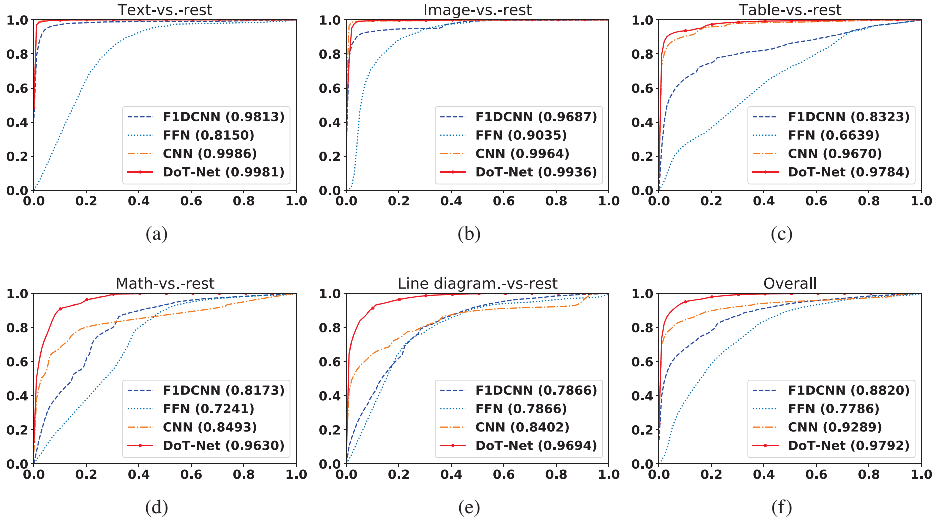
5) Conventional CNN: We also included the conventional CNN that is LeNet-5 [23], as a baseline method to compare with DoT-Net, although there is no study that directly used CNN for document layout classification. Three convolutional layers with 3×3 kernel size, 50 filers and TanH activations were used for optimal performance. Each convolutional layer was followed by max pooling layer with 2×2 pixel kernel. Dropout of 0.3 was used between max pooling and convolution layers for the optimization. A 100×100 2D image tile was introduced to the CNN model (Same as DoT-Net).

All benchmark methods, except CNN, were initially developed for the binary classifier of text vs. non-text blocks. However, the methods were easily extended for a multiclass classifier. Deep learning based methods (FFN, F1DCNN, HOGMLP, and CNN) were simply extended for multiclass classifiers by adding five nodes in the output layer with softmax activation. GSVM was also simply extended with

[1] We downloaded the dataset from: https://www.primaresearch.org/dataset.

TABLE I: Benchmark methods on the experimental settings

Input Classification	Tiles	Blocks
One-vs.-rest	FFN, F1DCNN, CNN, DoT-Net (Results in Table II)	GSVM, HOGMLP, FFN, F1DCNN, CNN, DoT-Net (Results in Table III)
Multiclass	FFN, F1DCNN, CNN, DoT-Net (Results in Table IV)	GSVM, HOGMLP, FFN, F1DCNN, CNN, DoT-Net (Results in Table V)



(a) (b) (c)

(d) (e) (f)

Fig. 4: ROC curves on tile-wise binary classification. (a) text-vs.-rest, (b) image-vs.-rest, (c) table-vs.-rest, (d) math-vs.-rest, (e) line diagram-vs.-rest, and (f) averaged overall ROC curves.

TABLE II: Performance with tile images in one-vs.-rest classification

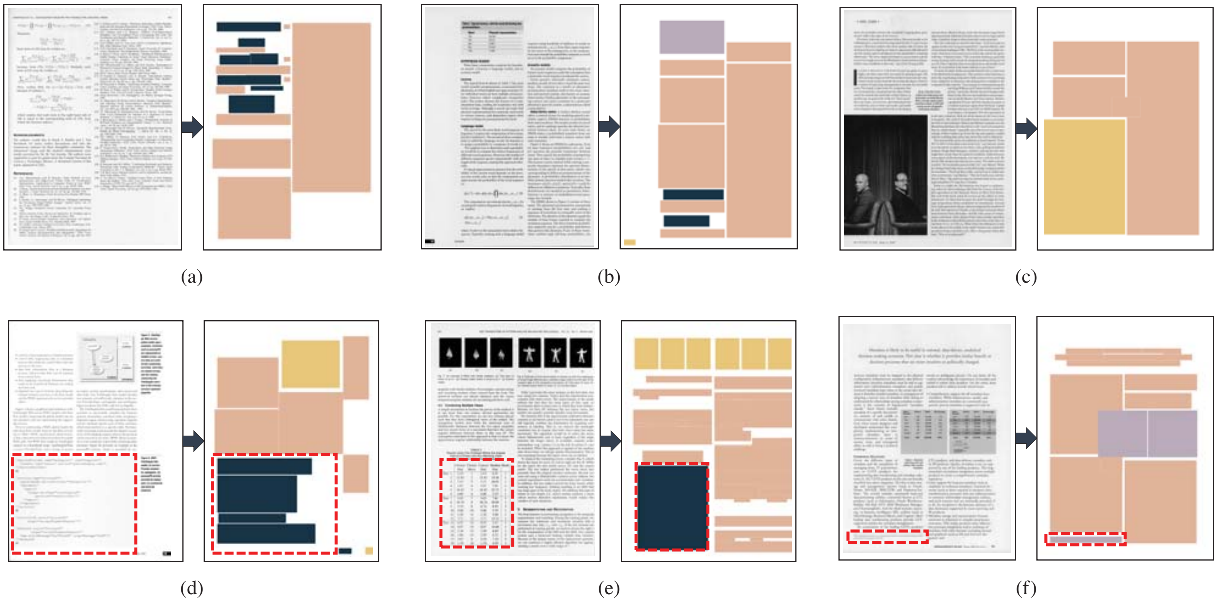
	Methods	Accuracy	F1 score	AUC
Text	F1DCNN	0.925 (0.047)	0.884 (0.052)	0.981 (0.014)
	FFN	0.774 (0.010)	0.801 (0.016)	0.815 (0.012)
	CNN	0.945 (0.016)	0.948 (0.014)	**0.998 (0.013)**
	DoT-Net	**0.981 (0.021)**	**0.982 (0.013)**	0.998 (0.015)
Image	F1DCNN	0.905 (0.036)	0.908 (0.041)	0.968 (0.030)
	FFN	0.841 (0.017)	0.850 (0.027)	0.903 (0.012)
	CNN	0.937 (0.009)	0.919 (0.011)	**0.994 (0.012)**
	DoT-Net	**0.970 (0.019)**	**0.971 (0.017)**	0.992 (0.018)
Table	F1DCNN	0.773 (0.046)	0.700 (0.051)	0.832 (0.066)
	FFN	0.596 (0.022)	0.644 (0.013)	0.663 (0.018)
	CNN	0.879 (0.028)	0.893 (0.020)	0.965 (0.038)
	DoT-Net	**0.917 (0.022)**	**0.919 (0.018)**	**0.978 (0.025)**
Math	F1DCNN	0.825 (0.053)	0.870 (0.037)	0.817 (0.041)
	FFN	0.705 (0.031)	0.748 (0.027)	0.724 (0.028)
	CNN	0.806 (0.015)	0.845 (0.013)	0.849 (0.017)
	DoT-Net	**0.900 (0.037)**	**0.898 (0.026)**	**0.963 (0.017)**
Line-diag.	F1DCNN	0.769 (0.027)	0.803 (0.052)	0.810 (0.033)
	FFN	0.737 (0.010)	0.753 (0.028)	0.7866 (0.012)
	CNN	0.769 (0.027)	0.718 (0.042)	0.840 (0.048)
	DoT-Net	**0.903 (0.024)**	**0.901 (0.027)**	**0.969 (0.010)**

Note: the average of the experiments repeated ten times and its standard errors in parenthesis are shown, and a bold-face represents the highest average on the experiments.

a conventional multiclass SVM model. The optimal hyper-parameters were determined by grid search with validation data for all of the benchmark methods.

For block-wise experiments, the tile-wise methods of FFN, F1DCNN, CNN, and DoT-Net first classified the tiles generated from a block, and then made the final decision by majority vote. The block-wise classifiers of GSVM and HOGMLP were not considered in the tile-wise experiments. The benchmark methods considered on each experiment setting are listed in Table I.

The experimental results of tile-wise and block-wise binary classification (one- vs. -rest) are summarized in Table II and Table III, respectively. For tile-wise binary classifications (Table II), DoT-Net obtained the highest accuracy and F1 score across the five classes. CNN achieved slightly higher AUCs (0.998 ± 0.013 and 0.994 ± 0.012) in the classes *text* and *image* than DoT-Net (0.998 ± 0.015 and 0.992 ± 0.018), while DoT-Net shows the outstanding AUCs in the rest classes (i.e. *table*, *math*, and *line-diagram*). In Fig. 4, the corresponding ROC curves demonstrate that the performances of CNN have dramatically dropped especially in the classes *math* and *line-diagram*, whereas DoT-Net shows an outstanding AUC with the five binary classifications on average. Similarly, DoT-Net outperformed the five benchmarks in most classes except *text* in the block-wise binary classification (see Table III). In the text-vs.-rest classification, CNN obtained the highest accuracy (0.981 ± 0.014) and AUC (0.997 ± 0.030), while DoT-Net obtained the highest F1 score (0.981 ± 0.022). In most cases, DoT-Net shows a robust predictive performance with the least standard errors.


(a)

(b)

(c)

(d)

(e)

(f)

Fig. 5: Examples of DoT-Net with sample documents. (a)–(c) Correctly classified blocks and (d)–(f) incorrectly classified blocks. Left side figures are the original documents and the right side figure illustrated the classification results with different colors (see color legend in Fig. 1).

More importantly, DoT-Net obtained the best performance among three measurements in the tile-wise multiclass classification in Table IV and block-wise multiclass classification in Table V. The proposed method Dot-Net appeared the best performance on the three measurements in both the tile-wise and block-wise multiclass classification. Especially, DoT-Net remarkably improved the model performance around 10% comparing to the second-ranked model, F1DCNN in the block-wise multiclass classification (see Table V). The results show that DoT-Net is a robust and accurate classifier for both binary and multiclass problems.

Furthermore, we applied DoT-Net for multiple documents which were not included in the training phase. We used the model with best F1-score. For document block detection, each page in a document was converted into a gray-scale image. Document blocks were detected by a modified bottom-up approach [24], and then tile images were generated by sliding a window of 100×100 pixels with stride of 30. Both correctly classified blocks (Fig. 5a–5c) and incorrectly classified blocks (Fig. 5d–5f) are shown, where left side figures are the original documents and the right side figure illustrated the classification results with different colors (see color legend in Fig. 1). In Fig. 5d, left bottom block (in red) of the document contains XML code. Dot-Net classified the block as *math*, whereas the ground truth was *text*. The block may be misclassified, due to the lack of enough numbers of training data of codes and its similar texture patterns with math (e.g., indentation and special character). In Fig. 5e, Dot-Net misclassified the *table* block in the left-bottom side of the document as *math*. The misclassification may be caused by the multiple numerical values that are not differentiated by lines in the table unlike most regular formatted tables have. In Fig. 5f, the block of *text* was classified as *table* due to the noise in the background.

IV. Conclusion

This paper presents a novel texture-based deep learning model (DoT-Net) for document layout classification. DoT-Net learns texture features by using dilated convolutional layers. Dilated convolutional layers followed by a max pooling layer enable one to capture texture features for a classification problem, whereas most dilated convolutional layers have been directly used as a deconvolutional layer. The experimental results with the public dataset demonstrated that our proposed method DoT-Net outperformed the current state-of-the-art methods in document layout classification.

Acknowledgments

This research was supported by GE Power. We would like to acknowledge the assistance of Dr. Jie Hao through the research.

References

[1] S. Bhowmik, R. Sarkar, M. Nasipuri, and D. Doermann, “Text and non-text separation in offline document images: a survey,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 21, no. 1, pp. 1–20, Jun 2018. [Online]. Available: https://doi.org/10.1007/s10032-018-0296-z

[2] A. Suvichakorn, S. Watcharabusaracum, and W. Sinthupinyo, “Simple layout segmentation of gray-scale document images,” in *International Workshop on Document Analysis Systems*. Springer, 2002, pp. 245–248.

[3] T. Guan and H. Zhu, “Atrous faster r-cnn for small scale object detection,” in *2017 2nd International Conference on Multimedia and Image Processing (ICMIP)*, March 2017, pp. 16–21.

TABLE III: Performance with block images in one-vs.-rest classification

	Methods	Accuracy	F1 score	AUC
Text	F1DCNN	0.936 (0.044)	0.941 (0.032)	0.940 (0.039)
	FFN	0.809 (0.021)	0.844 (0.028)	0.835 (0.043)
	CNN	**0.981 (0.014)**	0.976 (0.012)	**0.997 (0.030)**
	GSVM	0.821 (0.015)	0.831 (0.029)	0.815 (0.010)
	HOGMLP	0.783 (0.039)	0.762 (0.011)	0.790 (0.045)
	DoT-Net	0.978 (0.025)	**0.981 (0.022)**	0.991 (0.024)
Image	F1DCNN	0.912 (0.029)	0.905 (0.031)	0.941 (0.010)
	FFN	0.893 (0.021)	0.881 (0.024)	0.943 (0.032)
	CNN	0.943 (0.017)	0.932 (0.020)	0.966 (0.018)
	GSVM	0.869 (0.052)	0.862 (0.061)	0.872 (0.056)
	HOGMLP	0.851 (0.043)	0.825 (0.059)	0.863 (0.039)
	DoT-Net	**0.974 (0.017)**	**0.963 (0.027)**	**0.971 (0.019)**
Table	F1DCNN	0.813 (0.037)	0.845 (0.046)	0.872 (0.013)
	FFN	0.712 (0.017)	0.610 (0.026)	0.887 (0.032)
	CNN	0.843 (0.057)	0.829 (0.042)	0.861 (0.035)
	GSVM	0.581 (0.012)	0.771 (0.016)	0.482 (0.012)
	HOGMLP	0.682 (0.018)	0.773 (0.019)	0.635 (0.021)
	DoT-Net	**0.911 (0.028)**	**0.877 (0.019)**	**0.927 (0.038)**
Math	F1DCNN	0.748 (0.031)	0.763 (0.024)	0.817 (0.017)
	FFN	0.599 (0.033)	0.611 (0.021)	0.556 (0.033)
	CNN	0.612 (0.018)	0.561 (0.031)	0.646 (0.046)
	GSVM	0.633 (0.050)	0.549 (0.053)	0.650 (0.056)
	HOGMLP	0.689 (0.060)	0.662 (0.049)	0.699 (0.056)
	DoT-Net	**0.911 (0.040)**	**0.878 (0.032)**	**0.934 (0.038)**
Line-diagram	F1DCNN	0.828 (0.026)	0.794 (0.025)	0.855 (0.013)
	FFN	0.724 (0.027)	0.751 (0.018)	0.742 (0.018)
	CNN	0.639 (0.025)	0.582 (0.055)	0.674 (0.041)
	GSVM	0.735 (0.026)	0.742 (0.025)	0.749 (0.029)
	HOGMLP	0.756 (0.033)	0.781 (0.023)	0.747 (0.038)
	DoT-Net	**0.934 (0.013)**	**0.926 (0.019)**	**0.956 (0.025)**

Note: the average of the experiments repeated ten times and its standard errors in parenthesis are shown, and a bold-face represents the highest average on the experiments

TABLE IV: Performance with tile images in multiclass classification

	Accuracy	F1 score	AUC
F1DCNN	0.881 (0.022)	0.868 (0.028)	0.943 (0.024)
FFN	0.790 (0.046)	0.685 (0.017)	0.778 (0.043)
CNN	0.848 (0.027)	0.732 (0.043)	0.882 (0.016)
DoT-Net	**0.940 (0.019)**	**0.876 (0.009)**	**0.976 (0.012)**

Note: the average of the experiments repeated ten times and its standard errors in parenthesis are shown, and a bold-face represents the highest average on the experiments

TABLE V: Performance with block images in multiclass classification

	Accuracy	F1 score	AUC
F1DCNN	0.842 (0.019)	0.831 (0.013)	0.874 (0.023)
FFN	0.532 (0.035)	0.451 (0.027)	0.593 (0.041)
CNN	0.681 (0.024)	0.643 (0.013)	0.716 (0.029)
GSVM	0.449 (0.004)	0.394 (0.007)	0.512 (0.009
HOGMLP	0.491 (0.019)	0.371 (0.005)	0.532 (0.019)
DoT-Net	**0.941 (0.021)**	**0.929 (0.011)**	**0.952 (0.017)**

Note: the average of the experiments repeated ten times and its standard errors in parenthesis are shown, and a bold-face represents the highest average on the experiments

[4] Jin Wu, Wu-Mo Pan, Jian-Ming Jin, and Qing-Ren Wang, "Performance evaluation and benchmarking on document layout analysis algorithms," 2004.

[5] M. Diem, F. Kleber, and R. Sablatnig, "Text classification and document layout analysis of paper fragments," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2011.

[6] A. K. Sah, S. Bhowmik, S. Malakar, R. Sarkar, E. Kavallieratou, and N. Vasilopoulos, "Text and non-text recognition using modified hog descriptor," in *2017 IEEE Calcutta Conference (CALCON)*, Dec 2017, pp. 64–68.

[7] V. P. Le, N. Nayef, M. Visani, J. M. Ogier, and C. D. Tran, "Text and non-text segmentation based on connected component features," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2015.

[8] O. K. Oyedotun and A. Khashman, "Document segmentation using textural features summarization and feedforward neural network," *Applied Intelligence*, 2016.

[9] M. P. Viana and D. A. B. Oliveira, "Fast CNN-Based Document Layout Analysis," in *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2018.

[10] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for document image classification," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 3168–3172.

[11] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2015.

[12] E. Crestan and P. Pantel, "Web-scale knowledge extraction from semi-structured tables," 2010.

[13] E. Anisimova, P. Páta, and M. Blažek, "Stellar Object Detection Using the Wavelet Transform," *Acta Polytechnica*, vol. 51, no. 6, p. 9, 2011.

[14] A. Constantin, J. Ding, and Y. Lee, "Accurate road detection from satellite images using modified u-net," in *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Oct 2018, pp. 423–426.

[15] E. Anisimova, J. Bednář, and P. Páta, "Efficiency of wavelet coefficients thresholding techniques used for multimedia and astronomical image denoising," in *2013 International Conference on Applied Electronics*, Sep. 2013, pp. 1–4.

[16] Z. Hu, T. Turki, N. Phan, and J. T. L. Wang, "A 3d atrous convolutional long short-term memory network for background subtraction," *IEEE Access*, vol. 6, pp. 43 450–43 459, 2018.

[17] Y. Liu, X. Zhu, X. Zhao, and Y. Cao, "Adversarial learning for constrained image splicing detection and localization based on atrous convolution," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2019.

[18] N. Jin and Z. Long, "Effusion area segmentation for knee joint ultrasound image based on atrous-fcn with snake model algorithm," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Oct 2018, pp. 1–9.

[19] Z. Feng, H. Yong, and S. Xukun, "Granet: Global refinement atrous convolutional neural network for semantic scene segmentation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 1568–1572.

[20] Y. Liu and M. D. Levine, "Multi-path region-based convolutional neural network for accurate detection of unconstrained "hard faces"," in *2017 14th Conference on Computer and Robot Vision (CRV)*, May 2017, pp. 183–190.

[21] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, April 2018.

[22] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos, "ICDAR2009 page segmentation competition," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2009.

[23] N. Yi, C. Li, X. Feng, and M. Shi, "Research and improvement of convolutional neural network," in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, June 2018, pp. 637–640.

[24] V. Singh and B. Kumar, "Document layout analysis for Indian newspapers using contour based symbiotic approach," in *2014 International Conference on Computer Communication and Informatics: Ushering in Technologies of Tomorrow, Today, ICCCI 2014*, 2014.