IIIT-AR-13K: A New Dataset for Graphical Object Detection in Documents

Ajoy Mondal[1], Peter Lipps[2], and C V Jawahar[1]

[1] Centre for Visual Information Technology,
International Institute of Information Technology, Hyderabad, India
{ajoy.mondal,jawahar}@iiit.ac.in
[2] Open Text Software GmbH, Grasbrunn/Munich, Germany
peter.lipps@opentext.com

Abstract. We introduce a new dataset for graphical object detection in business documents, more specifically annual reports. This dataset, IIIT-AR-13K, is created by manually annotating the bounding boxes of graphical or page objects in publicly available annual reports. This dataset contains a total of 13K annotated page images with objects in five different popular categories — table, figure, natural image, logo, and signature. It is the largest manually annotated dataset for graphical object detection. Annual reports created in multiple languages for several years from various companies bring high diversity into this dataset. We benchmark IIIT-AR-13K dataset with two state of the art graphical object detection techniques using Faster R-CNN [20] and Mask R-CNN [11] and establish high baselines for further research. Our dataset is highly effective as training data for developing practical solutions for graphical object detection in both business documents and technical articles. By training with IIIT-AR-13K, we demonstrate the feasibility of a single solution that can report superior performance compared to the equivalent ones trained with a much larger amount of data, for table detection. We hope that our dataset helps in advancing the research for detecting various types of graphical objects in business documents[1].

Keywords: graphical object detection · annual reports · business documents · Faster R-CNN · Mask R-CNN.

1 Introduction

Graphical objects such as tables, figures, logos, equations, natural images, signatures, play an important role in the understanding of document images. They are also important for information retrieval from document images. Each of these graphical objects contains valuable information about the document in a compact form. Localizing such graphical objects is a primary step for understanding documents or information extraction/retrieval from the documents. Therefore, the detection of those graphical objects from the document images has attracted

[1] http://cvit.iiit.ac.in/usodi/iiitar13k.php

a lot of attention in the research community [7, 10, 13–16, 18, 21, 22, 24, 26, 27]. Large diversity within each category of the graphical objects makes detection task challenging. Researchers have explored a variety of algorithms for detecting graphical objects in the documents. Numerous benchmark datasets are also available in this domain to evaluate the performance of newly developed algorithms. In this paper, we introduce a new dataset, IIIT-AR-13K for graphical object detection.

Before we describe the details of the new dataset, we make the following observations:

1. State of the art algorithms (such as [7, 13, 14, 21, 22, 24, 26]) for graphical object detection are motivated by the success of object detection in computer vision (such as Faster R-CNN and Mask R-CNN). However, the high accuracy expectations in documents (similar to the high accuracy expectations in OCR) make the graphical object detection problem, demanding and thereby different compared to that of detecting objects in natural images.
2. With documents getting digitized extensively in many business workflows, automatic processing of business documents has received significant attention in recent years. However, most of the existing datasets for graphical object detection are still created from scientific documents such as technical papers.
3. In recent years, we have started to see large automatically annotated or synthetically created datasets for graphical object detection. We hypothesize that, in high accuracy regime, carefully curated small data adds more value than automatically generated large datasets. At least, in our limited setting, we validate this hypothesis later in this paper.

Popular datasets for graphical object detection (or more specifically table detection) are ICDAR 2013 table competition dataset [8] (i.e., ICDAR-2013), ICDAR 2017 competition on page object detection dataset [5] (i.e., ICDAR-POD-2017), cTDaR [6], UNLV [23], Marmot table recognition dataset [4], DeepFigures [25], PubLayNet [30], and TableBank [15]. Most of these existing datasets are limited with respect to their size (except DeepFigures, PubLayNet, and TableBank) or category labels (except ICDAR-POD-2017, DeepFigures, and PubLayNet). Most of them (except ICDAR-2013, cTDaR, and UNLV) consist of only scientific research articles, handwritten documents, and e-books. Therefore, they are limited in variations in layouts and structural variations in appearances.

On the contrary, business documents such as annual reports, pay slips, bills, receipt copies, etc. of the various companies are more heterogeneous to their layouts and complex visual appearance of the graphical objects. In such kind of heterogeneous documents, the detection of graphical objects becomes further difficult.

In this paper, we introduce a new dataset, named IIIT-AR-13K for localizing graphical objects in the annual reports (a specific type of business documents) of various companies. For this purpose, we randomly select publicly available annual reports in English and other languages (e.g., French, Japanese, Russian, etc.) of multiple (more than ten) years of twenty-nine different companies. We

manually annotate the bounding boxes of five different categories of graphical objects (i.e., table, figure, natural image, logo, and signature), which frequently appear in the annual reports. IIIT-AR-13K dataset contains 13K annotated pages with 16K tables, 3K figures, 3K natural images, 0.5K logos, and 0.6K signatures. To the best of the author's knowledge, the newly created dataset is the largest among all the existing datasets where ground truths are annotated manually for detecting graphical objects (more than one category) in documents.

We use Faster R-CNN [20] and Mask R-CNN [11] algorithms to benchmark the newly created dataset for graphical object detection task in business documents. Experimentally, we observe that the creation of a model trained with IIIT-AR-13K dataset achieves better performance than the model trained with the larger existing datasets. From the experiments, we also observe that the model trained with the larger existing datasets achieves the best performance by fine-tuning with a small number (only 1K) of images from IIIT-AR-13K dataset.

Our major contributions/claims are summarized as follows:

- We introduce a highly variate new dataset for localizing graphical objects in documents. The newly created dataset is the largest among the existing datasets where ground truth is manually annotated for the graphical object detection task. It also has a larger label space compared to most existing datasets.
- We establish Faster R-CNN and Mask R-CNN based benchmarks on the popular datasets. We report very high quantitative detection rates on the popular benchmarks.
- Though smaller than some of the recent datasets, this dataset is more effective as training data for the detection task due to the inherent variations in the object categories. This is empirically established by creating a unique model trained with IIIT-AR-13K dataset to detect graphical objects in all existing datasets.
- Models trained with the larger existing datasets achieve the best performance after fine-tuning with a very limited number (only 1K) of images from IIIT-AR-13K dataset.

2 Preliminaries

Graphical objects such as tables, various types of figures (e.g., bar chart, pie chart, line plot, natural image, etc.) equations, and logos in documents contain valuable information in a compact form. Understanding of document images requires localizing of such graphical objects as an initial step. Several datasets (e.g., ICDAR-2013 [8], ICDAR-POD-2017 [5], cTDaR [6], UNLV [23], Marmot [4], DeepFigures [25], PubLayNet [30], and TableBank [15]) exist in the literature, which are dedicated to localize graphical objects (more specifically tables) in the document images. Ground truth bounding boxes of ICDAR-2013, ICDAR-POD-2017, cTDaR, UNLV, and Marmot are annotated manually. Ground truth bounding boxes of DeepFigures, PubLayNet, and TableBank are generated automatically. Among these datasets, only ICDAR-POD-2017, DeepFigures, and

PubLayNet are aimed to address localizing a wider class of graphical objects. Other remaining datasets are designed only for localizing tables. Some of these datasets (e.g., ICDAR-2013, cTDaR, UNLV, and TableBank) are also used for table structure recognition and table recognition tasks in the literature. Some other existing datasets such as SciTSR [2], Table2Latex [3], and PubTabNet [29] are used exclusively for table structure recognition and table recognition purpose in the literature.

2.1 Related Datasets

ICDAR-2013 [8]: This dataset is one of the most cited datasets for the task of table detection and table structure recognition. It contains 67 PDFs which corresponds to 238 page images. Among them, 40 PDFs are taken from the US Government and 27 PDFs from EU. Among 238 page images, only 135 page images contain tables, in total 150 tables. This dataset is popularly used to evaluate the algorithms for both table detection and structure recognition task.

ICDAR-POD-2017 [5]: It focuses on the detection of various graphical objects (e.g., tables, equations, and figures) in the document images. It is created by annotating 2417 document images selected from 1500 scientific papers of CiteSeer. It includes a large variety in page layout - single-column, double-column, multi-column, and a significant variation in the object structure. This dataset is divided into (i) training set consisting of 1600 images, and (ii) test set comprising 817 images.

cTDaR [6]: This dataset consists of modern and archival documents with various formats, including document images and born-digital formats such as PDF. The images show a great variety of tables from hand-drawn accounting books to stock exchange lists and train timetable, from record books of prisoner lists, tables from printed books, production census, etc. The modern documents consist of scientific journals, forms, financial statements, etc. Annotations correspond to table regions, and cell regions are available. This dataset contains (i) training set - consists of 600 annotated archival and 600 annotated modern document images with table bounding boxes, and (ii) test set - includes 199 annotated archival and 240 annotated modern document images with table bounding boxes.

Marmot [4]: It consists of 2000 Chinese and English pages at the proportion of about 1:1. Chinese pages are selected from over 120 e-Books with diverse subject areas provided by Founder Apabi library. Not more than 15 pages are selected from each Book. English pages are selected over 1500 journal and conference papers published during 1970-2011 from the Citeseer website. The pages show a great variety in layout - one-column and two-column, language type, and table styles. This dataset is used for both table detection and structure recognition tasks.

UNLV [23]: It contains 2889 document images of various categories: technical reports, magazines, business letters, newspapers, etc. Among them, only 427 images include 558 table zones. Annotations for table regions and cell regions are available. This dataset is also used for both table detection and table structure recognition tasks.

DeepFigures [25]: Siegel *et al.* [25] create this dataset by automatically annotating pages of two large web collections of scientific documents (arXiv and PubMed). This dataset consists of 5.5 million pages with 1.4 million tables and 4.0 million figures. It can be used for detecting tables and figures in the document images.

PubLayNet [30]: *and* ***PubTabNet [29]*** PubLayNet consists of 360K page images with annotation of various layout elements such as text, title, list, figure, and table. It is created by automatically annotating 1 million PubMed Central^{TM} PDF articles. It contains 113K table regions and 126K figure regions in total. It is mainly used for layout analysis purposes. However, it can also be used for table and figure detection tasks. PubTabNet is the largest dataset for the table recognition task. It consists of 568K images of heterogeneous tables extracted from scientific articles (in PDF format). Ground truth corresponds to each table image represents structure information and text content of each cell in HTML format.

TableBank [15]: This dataset contains 417K high-quality documents with tables. This dataset is created by downloading LaTex source code from the year 2014 to the year 2018 through bulk data access in arXiv. It consists of 163K word document, 253K LaTex document and in total 417K document images with annotated table regions. Structure level annotation is available for 57K word documents, 88K LaTex documents and in total 145K document images.

2.2 Related Work in Graphical Object Detection

Localizing graphical objects (e.g., tables, figures, mathematical equations, logos, signatures, etc.) is the primary step for understanding any documents. Recent advances in object detection in natural scene images using deep learning inspire researchers [7, 13, 14, 16, 18, 21, 22, 24, 26] to develop deep learning based algorithms for detecting graphical objects in documents. In this regards, various researchers like Gilani *et al.* [7], Schreiber *et al.* [22], Siddiqui *et al.* [24] and Sun *et al.* [26] employ Faster R-CNN [20] model for detecting table in documents. In [13], the authors use YOLO [19] to detect tables in documents. Li *et al.* [16] use Generative Adversarial Networks (GAN) [9] to extract layout feature to improve table detection accuracy.

Few researchers [14,21] focus on detection of various kinds of graphical objects (not just tables). Kavasidis *et al.* [14] propose saliency based technique to detect three types of graphical objects — table, figure and mathematical equation. In [21], Mask R-CNN is explored to detect various graphical objects.

2.3 Baseline Methods for Graphical Object Detection

We use Faster R-CNN [20] and Mask R-CNN [11] as baselines for detecting graphical objects in annual reports. We use publicly available implementations of Faster R-CNN [28] and Mask R-CNN [1] for this experiment. We train both the models with the images of training set of IIIT-AR-13K dataset for the empirical studies.

Faster R-CNN: The model is implemented using PyTorch; trained and evaluated on NVIDIA TITAN X GPU with 12GB memory with batch size of 4. The input images are resized to a fixed size of 800 $\times$ 1024 by preserving original aspect ratio. We use the pre-trained ResNet-101 [12] backbone on MS-COCO [17] dataset. We use five different anchor scales (i.e., 32, 64, 128, 256, and 512) and five anchor ratios (i.e., 1, 2, 3, 4, and 5) so that the region proposals can cover almost every part of the image irrespective of the image size. We use stochastic gradient descent (SGD) as an optimizer with initial learning rate = 0.001 and the learning rate decays after every 5 epochs and it is equal to 0.1 times of the previous value.

Mask R-CNN: Every input image is rescaled to a fixed size of 800 $\times$ 1024 preserving the original aspect ratio. We use the pre-trained ResNet-101 [12] backbone on MS-COCO [17] dataset. We use Tensorflow/Keras for the implementation, and train and evaluate on NVIDIA TITAN X GPU that has 12GB memory with a batch size of 1. We further use 0.5, 1, and 2 as the anchor scale values and anchor box sizes of 32, 64, 128, 256, and 512. Further, we train a total of 80 epochs. We train all FPN and subsequent layers for first 20 epochs, the next 20 epochs for training FPN + last 4 layers of ResNet-101; and last 40 epochs for training all layers of the model. During the training process, we use 0.001 as the learning rate, 0.9 as the momentum and 0.0001 as the weight decay.

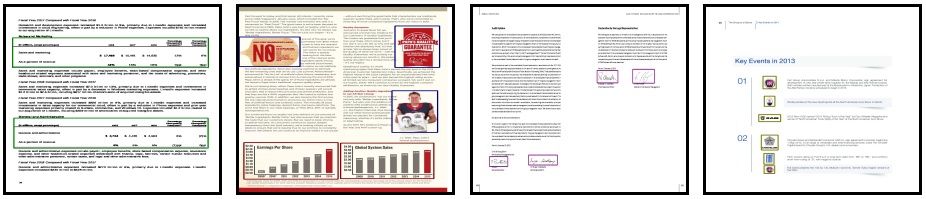


Fig. 1: Sample annotated document images of IIIT-AR-13K dataset. **Dark Green:** indicates ground truth bounding box of table, **Dark Red:** indicates ground truth bounding box of figure, **Dark Blue:** indicates ground truth bounding box of natural image, **Dark Yellow:** indicates ground truth bounding box of logo and **Dark Pink:** indicates ground truth bounding box of signature.

3 IIIT-AR-13K Dataset

3.1 Details of the Dataset

For detecting various types of graphical objects (e.g., table, figure, natural image, logo, and signature) in the business documents, we generate a new dataset, named IIIT-AR-13K. The newly generated dataset consists of 13K pages of publicly available annual reports. Annual reports in English and non-English languages (e.g., French, Japanese, Russian, etc.) of multiple (more than ten) years of twenty-nine different companies. Annual reports contain various types of graphical objects such as tables, various types of graphs (e.g., bar chart, pie chart, line plot, etc.), images, companies' logos, signatures, stamps, sketches, etc. However, we only consider five categories of graphical objects, e.g., table, figure (including both graphs and sketches), natural image (including images), logo, and signature.

Object Category	**IIIT-AR-13K**			
	Training	**Validation**	**Test**	**Total**
Page	$9333 \approx 9K$	$1955 \approx 2K$	$2120 \approx 2K$	$13415 \approx 13K$
Table	$11163 \approx 11K$	$2222 \approx 2K$	$2596 \approx 3K$	$15981 \approx 16K$
Figure	$2004 \approx 2K$	$481 \approx 0.4K$	$463 \approx 0.4K$	$2948 \approx 3K$
Natural Image	$1987 \approx 2K$	$438 \approx 0.4K$	$455 \approx 0.4K$	$2880 \approx 3K$
Logo	$379 \approx 0.3K$	$67 \approx 0.06K$	$135 \approx 0.1K$	$581 \approx 0.5K$
Signature	$420 \approx 0.4K$	$108 \approx 0.9K$	$92 \approx 0.09K$	$620 \approx 0.6K$

Table 1: Statistics of our newly generated IIIT-AR-13K dataset.

Annual reports in English and other languages of more than ten years of twenty-nine different companies are selected to increase diversity with respect to the language, layout, and each category of graphical objects. We manually annotate the bounding boxes of each type of graphical object. Figure 1 shows some sample bounding box annotated images. Finally, we have 13K annotated pages with 16K tables, 3K figures, 3K natural images, 0.5K logos, and 0.6K signatures. The dataset is divided into (i) training set consisting of 9K document images, (ii) validation set containing 2K document images, and (iii) test set consists of 2K document images. For every company, we randomly choose 70%, 15%, and remaining 15% of the total pages as training, validation, and test, respectively. Table 1 shows the statistics of the newly generated dataset.

3.2 Comparison with the Existing Datasets

Table 2 presents the comparison of our dataset with the existing datasets. From the table, it is clear that with respect to label space, all the existing datasets (except ICDAR-POD-2017, DeepFigures, and PubLayNet) containing only one object category, i.e., table, are subsets of our newly generated IIIT-AR-13K dataset

(containing five object categories). Most of the existing datasets (except ICDAR-2013, cTDaR, and UNLV) consist of only scientific research articles, handwritten documents, and e-books. The newly generated dataset includes annual reports, one specific type of business document. It is the largest among the existing datasets where ground truths are annotated manually for the detection task. It consists of a large variety of pages of annual reports compared to scientific articles of existing datasets.

Dataset	**Category Label**	**#Images**	**Task**				**#Tables**
			POD	**TD**	**TSR**	**TR**	
ICDAR-2013 [8]	1: T	238	✓	✓	✓	✓	150
ICDAR-POD-2017 [5]	3: T, F, E	2417	✓	✓	×	×	1020
cTDaR [6]	1: T	2K	✓	✓	✓	✓	3.6K
Marmot [4]	1: T	2K	✓	✓	×	×	958
UNLV [23]	1: T	427	✓	✓	✓	✓	558
IIIT-AR-13k	**5: T, F, NI, L, S**	**13K**	✓	✓	×	×	**16K**
DeepFigures [25][2]	2: T, F	5.5M	✓	✓	×	×	1.4M
PubLayNet [30][2]	5: T, F, TL, TT, LT	360K	✓	✓	×	×	113K
TableBank [15][2]	1: T	417K	✓	✓	×	×	417K
			×	×	✓	×	145K
SciTSR [2][3]	1: T	-	×	×	✓	✓	15K
Table2Latex [3][3]	1: T	-	×	×	✓	✓	450K
PubTabNet [29][3]	1: T	-	×	×	✓	✓	568K

Table 2: Statistics of existing datasets along with newly generated dataset for graphical object detection task in document images. **T:** indicates table. **F:** indicates figure. **E:** indicates equation. **NI:** indicates natural image. **L:** indicates logo. **S:** indicates signature. **TL:** indicates title. **TT:** indicates text. **LT:** indicates list. **POD:** indicates page object detection. **TD:** indicates table detection. **TSR:** indicates table structure recognition. **TR:** indicates table recognition.

3.3 Diversity in Object Category

We select annual reports (more than ten years) of twenty-nine different companies for creating this dataset. Due to the consideration of annual reports of several years of various companies, there is a significant variation in layouts (e.g., single-column, double-column, and triple-column) and graphical objects like tables, figures, natural images, logos, and signatures. The selected documents are heterogeneous. Heterogeneity increases the variability within each object category and also increases the similarity between object categories. A significant variation in layout structure, the similarity between object categories, and diversity within object categories make this dataset more complex for graphical detection tasks. Figure 2 illustrates the significant variations in table structures and figures in the newly generated dataset.

[2] Ground truth bounding boxes are annotated automatically.
[3] Dataset is dedicated only for table structure recognition and table recognition.

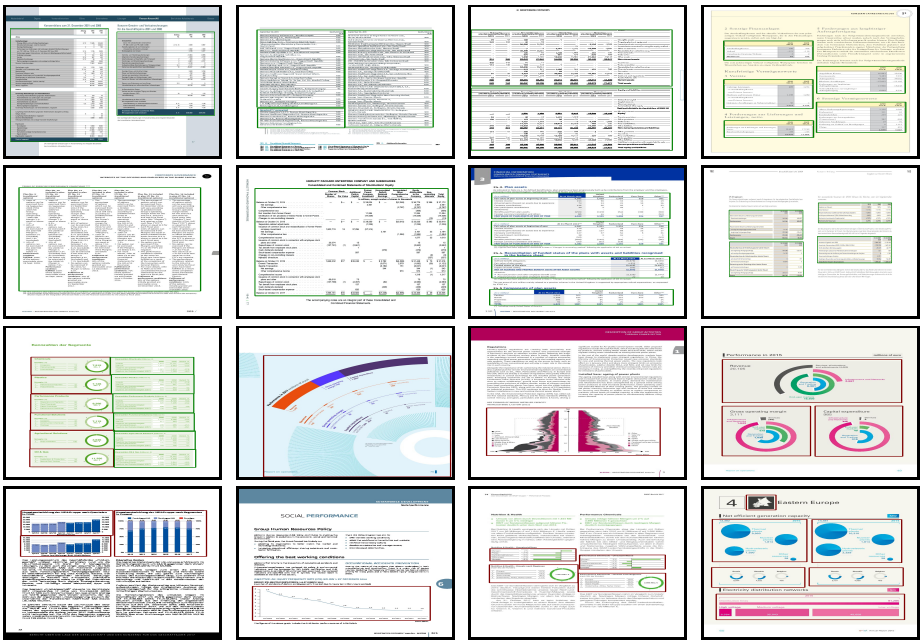


Fig. 2: Sample annotated images with large variation in tables and figures.

3.4 Performance for Detection using Baseline Methods

Quantitative results of detection on the validation and test sets of IIIT-AR-13K dataset using baseline approaches - Faster R-CNN and Mask R-CNN are summarized in Table 3. From the table, it is observed that Mask R-CNN produces better results than Faster R-CNN. Among all the object categories, both the baselines obtain the best results for table and worst results for logo. This is because of highly imbalanced training set (11K tables and 0.3K logos).

3.5 Effectiveness of IIIT-AR-13K Over Existing Larger Datasets

The newly generated dataset IIIT-AR-13K is smaller (with respect to number of document images) than some of the existing datasets (e.g., DeepFigure, TableBank, and PubLayNet) for the graphical object (i.e., table) detection task. We establish the effectiveness of the smaller IIIT-AR-13K dataset over the larger existing datasets - DeepFigure, TableBank, and PubLayNet for graphical object (i.e., table common object category of all datasets) detection task. In [15], the authors experimentally show that TableBank dataset is more effective than the largest DeepFigure dataset for table detection. In this paper, we use TableBank and PubLayNet datasets to establish the effectiveness of the newly generated dataset over the existing datasets for table detection. For this purpose, Mask

Dataset	Category	Faster R-CNN				Mask R-CNN			
		R↑	P↑	F↑	mAP↑	R↑	P↑	F↑	mAP↑
Validation	Table	0.9571	0.9260	0.9416	0.9554	**0.9824**	**0.9664**	**0.9744**	**0.9761**
	Figure	0.8607	0.7800	0.8204	0.8103	**0.8699**	**0.8326**	**0.8512**	**0.8391**
	Natural Image	0.9027	0.8607	0.8817	0.8803	**0.9461**	**0.8820**	**0.9141**	**0.9174**
	Logo	0.8566	0.4063	0.6315	0.6217	**0.8852**	**0.4122**	**0.6487**	**0.6434**
	Signature	0.9411	0.8000	0.8705	0.9135	**0.9633**	**0.8400**	**0.9016**	**0.9391**
	Average	0.9026	0.7546	0.8291	0.8362	**0.9294**	**0.7866**	**0.8580**	**0.8630**
Test	Table	0.9512	0.9234	0.9373	0.9392	**0.9711**	**0.9715**	**0.9713**	**0.9654**
	Figure	0.8530	0.7582	0.8056	0.8332	**0.8898**	**0.7872**	**0.8385**	**0.8686**
	Natural Image	0.8745	0.8631	0.8688	0.8445	**0.9179**	**0.8625**	**0.8902**	**0.8945**
	Logo	0.5933	0.3606	0.4769	0.4330	**0.6330**	**0.3920**	**0.5125**	**0.4699**
	Signature	0.8868	0.7753	0.8310	0.8981	**0.9175**	**0.7876**	**0.8525**	**0.9115**
	Average	0.8318	0.7361	0.7839	0.7896	**0.8659**	**0.7601**	**0.8130**	**0.8220**

Table 3: Quantitative results of two baseline approaches for detecting graphical objects. **R:** indicates Recall. **P:** indicates Precision. **F:** indicates F-measure. **mAP:** indicates mean average precision.

R-CNN model is trained with training set of each of three datasets - TableBank, PubLayNet, and IIIT-AR-13K. These trained models are individually evaluated on test set of existing datasets - ICDAR-2013, ICDAR-POD-2017, cTDaR, UNLV, Marmot, PubLayNet. We name these experiments as

(i) **Experiment-I:** Mask R-CNN trained with TableBank (LaTex) dataset.
(ii) **Experiment-II:** Mask R-CNN trained with TableBank (Word) dataset.
(iii) **Experiment-III:** Mask R-CNN trained with TableBank (LaTex+Word) dataset.
(iv) **Experiment-IV:** Mask R-CNN trained with PubLayNet dataset.
(v) **Experiment-V:** Mask R-CNN trained with IIIT-AR-13K dataset.

Observation-I - Without Fine-tuning: To establish effectiveness of the newly generated IIIT-AR-13K dataset over larger existing datasets, we do experiments: Experiment-I, Experiment-II, Experiment-III, Experiment-IV, and Experiment-V. Table 4 illustrates quantitative results of these experiments. The table highlights that Experiment-V produces the best results (with respect to F-measure and mAP) for cTDaR, UNLV, and IIIT-AR-13K datasets. Though Experiment-I produces the best mAP value for Marmot dataset, the performance of Experiment-V on this dataset is very close to the performance of Experiment-I. For ICDAR-2017 dataset, the performance of Experiment-V is lower than Experiment-I and Experiment-III with respect to F-measure. The table highlights that, even though IIIT-AR-13K is significantly smaller than TableBank and PubLayNet, it is more effective for training a model for detecting tables in the existing datasets.

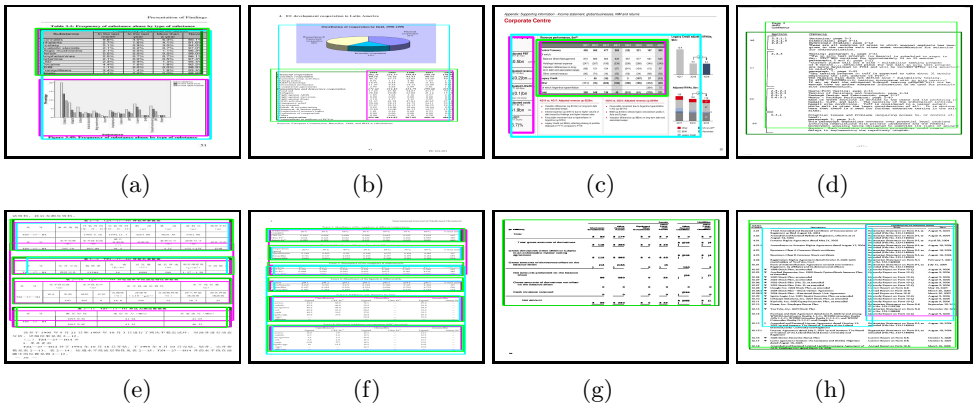
Figure 3 shows the predicted bounding boxes of tables in the pages of several datasets using three different models. Here, dark Green, Pink, Cyan, and Light

Test Dataset	Training Dataset	Quantitative Score			
		R↑	**P↑**	**F↑**	**mAP↑**
ICDAR-2013	TBL	0.9454	0.9341	**0.9397**	0.9264
	TBW	0.9090	**0.9433**	0.9262	0.8915
	TBLW	0.9333	0.9277	0.9305	0.9174
	PubLayNet	0.9272	0.8742	0.9007	0.9095
	IIIT-AR-13K	**0.9575**	0.8977	0.9276	**0.9393**
ICDAR-2017	TBL	0.9274	0.7016	0.8145	0.8969
	TBW	0.8107	0.5908	0.7007	0.7520
	TBLW	**0.9369**	**0.7406**	**0.8387**	**0.9035**
	PubLayNet	0.8296	0.6368	0.7332	0.7930
	IIIT-AR-13K	0.8675	0.6311	0.7493	0.7509
cTDaR	TBL	0.7006	0.7208	0.71078	0.5486
	TBW	0.6405	0.7582	0.6994	0.5628
	TBLW	0.7230	0.7481	0.7356	0.5922
	PubLayNet	0.6391	0.7231	0.6811	0.5610
	IIIT-AR-13K	**0.8097**	**0.8224**	**0.8161**	**0.7478**
UNLV	TBL	0.5806	0.6612	0.6209	0.5002
	TBW	0.5035	0.8005	0.6520	0.4491
	TBLW	0.6362	0.6787	0.6574	0.5513
	PubLayNet	0.7329	0.8329	0.7829	0.6950
	IIIT-AR-13K	**0.8602**	**0.7843**	**0.8222**	**0.7996**
Marmot	TBL	0.8860	**0.8840**	0.8850	**0.8465**
	TBW	0.8423	0.8808	0.8615	0.8026
	TBLW	**0.8919**	0.8802	**0.8860**	0.8403
	PubLayNet	0.8549	0.8214	0.8382	0.7987
	IIIT-AR-13K	0.8852	0.8075	0.8464	0.8464

Table 4: Performance of table detection in the existing datasets. Model is trained with only training images containing tables of the respective datasets. **TBL:** indicates TableBank (LaTex). **TBW:** indicates TableBank (Word). **TBLW:** indicates TableBank (LaTex+Word). **R:** indicates Recall. **P:** indicates Precision. **F:** indicates F-measure. **mAP:** indicates mean average precision.

Green colored rectangles highlighted the ground truth and predicted bounding boxes of tables using Experiment-I, Experiment-IV, and Experiment-V, respectively. In those images, Experiment-V correctly detects the tables. At the same time, either Experiment-I or Experiment-IV or Experiment-I and Experiment-IV fails to predict the bounding boxes of the tables accurately.

Observation-II - Fine-tuning with Complete Training Set: From Table 5, it is also observed that the performance of Experiment-I, Experiment-II, Experiment-III, and Experiment-IV on the validation set and test set of IIIT-AR-13K dataset is significantly worse (10% in case of F-measure and 15% in case of mAP) than the performance of Experiment-V. On the contrary, when we fine-tune with the training set of IIIT-AR-13K dataset, all the fine-tuning ex-



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3: Few examples of predicted bounding boxes of tables in various datasets - (a) ICDAR-2013, (b) ICADR-POD-2017, (c) cTDaR, (d) UNLV, (e) Marmot, (f) PubLayNet, (g) IIIT-AR-13K (validation), (h) IIIT-AR-13K (test) using **Experiment-I**, **Experiment-IV** and **Experiment-V**. **Experiment-I:** Mask R-CNN trained with TableBank (LaTex) dataset. **Experiment-IV:** Mask R-CNN trained with PubLayNet dataset. **Experiment-V:** Mask R-CNN trained with IIIT-AR-13K dataset. **Dark Green:** rectangle highlights the ground truth bounding boxes of tables. **Pink, Cyan and Light Green:** rectangles indicate the predicted bounding boxes of tables using **Experiment-I**, **Experiment-IV** and **Experiment-V**, respectively.

periments obtain similar outputs compared to Experiment-V. In the case of the validation set of PubLayNet dataset, Experiment-IV outperforms Experiment-II and Experiment-V; and the performances of Experiment-I and Experiment-III are very close to the performance of Experiment-IV. When we fine-tune with the training set of PubLayNet dataset, the fine-tuned model achieves similar performance compared to Experiment-IV. These experiments highlight that fine-tuning with the complete training set is effective for both the larger existing datasets and as well as IIIT-AR-13K dataset.

Observation-III - Fine-tuning with Partial Training Set: From Table 5, we observe that the performance of the trained model using TableBank and PubLayNet on validation and test sets of IIIT-AR-13K dataset is (10-20%) less than the performance of model trained with training set of IIIT-AR-13K. When we fine-tune the models with a partial training set of only 1K randomly selected images of IIIT-AR-13K, the models also achieve similar outputs (see Table 5). On the other hand, the performance of the trained model with IIIT-AR-13K dataset on validation set of PubLayNet is (10-15%) is less than models trained with TableBank and PubLayNet. When we fine-tune the model (trained with IIIT-AR-13K dataset) with the complete training set of PubLayNet, the model obtains very close output to the models trained with TableBank and PubLayNet. But when we fine-tune the model (trained with IIIT-AR-13K dataset) with the

Test Dataset	Training Dataset	Fine-tune	Quantitative Score			
			R↑	P↑	F↑	mAP↑
PubLayNet (validation)	TBL		0.9859	0.6988	0.8424	0.9461
	TBW		0.9056	0.7194	0.8125	0.8276
	TBLW		0.9863	0.7258	0.8560	0.9557
	PubLayNet		**0.9886**	0.7780	0.8833	0.9776
	IIIT-AR-13K		0.9555	0.5530	0.7542	0.8696
	IIIT-AR-13K	PubLayNet(F-TRS)	0.9882	**0.7898**	**0.8890**	**0.9781**
	IIIT-AR-13K	PubLayNet(P-TRS)	0.9869	0.4653	0.7261	0.9712
IIIT-AR-13K (validation)	TBL		0.8109	0.7370	0.7739	0.7533
	TBW		0.7641	0.8214	0.7928	0.7230
	TBLW		0.8217	0.7345	0.7781	0.7453
	PubLayNet		0.8100	0.7092	0.7596	0.7382
	IIIT-AR-13K		**0.9891**	0.7998	0.8945	0.9764
	TBL	IIIT-AR-13K(F-TRS)	0.9869	0.8600	0.9234	0.9766
	TBW		0.9815	0.8065	0.8940	0.9639
	TBLW		0.9873	0.8614	0.9244	**0.9785**
	PubLayNet		0.9842	0.8344	0.9093	0.9734
	TBL	IIIT-AR-13K(P-TRS)	0.9747	0.8790	0.9269	0.9648
	TBW		0.9734	0.8767	0.9251	0.9625
	TBLW		0.9783	**0.8998**	**0.9391**	0.9717
	PubLayNet		0.9797	0.8732	0.9264	0.9687
IIIT-AR-13K (test)	TBL		0.8023	0.7428	0.7726	0.7400
	TBW		0.7704	0.8396	0.8050	0.7276
	TBLW		0.8278	0.7500	0.7889	0.7607
	PubLayNet		0.8093	0.7302	0.7697	0.7313
	IIIT-AR-13K		**0.9826**	0.8361	0.9093	0.9688
	TBL	IIIT-AR-13K(F-TRS)	0.9761	0.8774	0.9267	0.9659
	TBW		0.9753	0.8239	0.8996	0.9596
	TBLW		0.9776	0.8788	0.9282	**0.9694**
	PubLayNet		0.9753	0.8571	0.9162	0.9642
	TBL	IIIT-AR-13K(P-TRS)	0.9637	0.9022	0.9330	0.9525
	TBW		0.9699	0.8922	0.9311	0.9615
	TBLW		0.9718	**0.9091**	**0.9405**	0.9629
	PubLayNet		0.9684	0.8905	0.9294	0.9552

Table 5: Performance of table detection using fine-tuning. **TBL:** indicates TableBank (LaTex). **TBW:** indicates TableBank (Word). **TBLW:** indicates TableBank (LaTex+Word). **F-TRS:** indicates complete training set. **P-TRS:** indicates partial training set i.e., only 1K randomly selected training images. **R:** indicates Recall. **P:** indicates Precision. **F:** indicates F-measure. **mAP:** indicates mean average precision.

partial training set (only 1K images) of PubLayNet, the model is unable to obtain similar output (with respect to F-measure) to the models trained with TableBank and PubLayNet. These experiments also highlight that the newly created IIIT-AR-13K dataset is more effective for fine-tuning.

4 Summary and Observations

This paper presents a new dataset IIIT-AR-13K for detecting graphical objects in business documents, specifically annual reports. It consists of 13K pages of annual reports of more than ten years of twenty-nine different companies with bounding box annotation of five different object categories — table, figure, natural image, logo, and signature.

- The newly generated dataset is the largest manually annotated dataset for graphical object detection purpose, at this stage.
- This dataset has more labels and diversity compared to most of the existing datasets.
- Though IIIT-AR-13K is smaller than the existing automatic annotated datasets — DeepFigures, PubLayNet, and TableNet, the model trained with IIIT-AR-13K performs better than the model trained with larger datasets for detecting tables in document images in most cases.
- Models trained with the existing datasets also achieve better performance by fine-tuning with a limited number of training images from IIIT-AR-13K.

We believe that this dataset will aid research in detecting tables and other graphical objects in business documents.

References

1. Abdulla, W.: Mask R-CNN for object detection and instance segmentation on Keras and Tensorflow. GitHub repository (2017)
2. Chi, Z., Huang, H., Xu, H.D., Yu, H., Yin, W., Mao, X.L.: Complicated table structure recognition. arXiv (2019)
3. Deng, Y., Rosenberg, D., Mann, G.: Challenges in end-to-end neural scientific table recognition. In: ICDAR (2019)
4. Fang, J., Tao, X., Tang, Z., Qiu, R., Liu, Y.: Dataset, ground-truth and performance metrics for table detection evaluation. In: WDAS (2012)
5. Gao, L., Yi, X., Jiang, Z., Hao, L., Tang, Z.: ICDAR 2017 competition on page object detection. In: ICDAR (2017)
6. Gao, L., Djean, H., Yan, Q., Kleber, F., Huang, Y., Meunier, J.L., Fang, Y.: ICDAR 2019 competition on table detection and recognition (cTDaR). In: ICDAR (2019)
7. Gilani, A., Qasim, S.R., Malik, I., Shafait, F.: Table detection using deep learning. In: ICDAR (2017)
8. Göbel, M., Hassan, T., Oro, E., Orsi, G.: ICDAR 2013 table competition. In: ICDAR (2013)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
10. Hao, L., Gao, L., Yi, X., Tang, Z.: A table detection method for PDF documents based on convolutional neural networks. In: Workshop on DAS (2016)

11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
13. Huang, Y., Yan, Q., Li, Y., Chen, Y., Wang, X., Gao, L., Tang, Z.: A YOLO-based table detection method. In: ICDAR (2019)
14. Kavasidis, I., Pino, C., Palazzo, S., Rundo, F., Giordano, D., Messina, P., Spampinato, C.: A saliency-based convolutional neural network for table and chart detection in digitized documents. In: International Conference on Image Analysis and Processing (2019)
15. Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., Li, Z.: TableBank: Table benchmark for image-based table detection and recognition. In: ICDAR (2019)
16. Li, Y., Yan, Q., Huang, Y., Gao, L., Tang, Z.: A GAN-based feature generator for table detection. In: ICDAR (2019)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
18. Melinda, L., Bhagvati, C.: Parameter-free table detection method. In: ICDAR (2019)
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015)
21. Saha, R., Mondal, A., Jawahar, C.V.: Graphical object detection in document images. In: ICDAR (2019)
22. Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In: ICDAR (2017)
23. Shahab, A., Shafait, F., Kieninger, T., Dengel, A.: An open approach towards the benchmarking of table structure recognition systems. In: DAS (2010)
24. Siddiqui, S.A., Malik, M.I., Agne, S., Dengel, A., Ahmed, S.: DeCNT: Deep deformable CNN for table detection. IEEE Access (2018)
25. Siegel, N., Lourie, N., Power, R., Ammar, W.: Extracting scientific figures with distantly supervised neural networks. In: ACM/IEEE on joint conference on digital libraries (2018)
26. Sun, N., Zhu, Y., Hu, X.: Faster R-CNN based table detection combining corner locating. In: ICDAR (2019)
27. Tran, D.N., Tran, T.A., Oh, A., Kim, S.H., Na, I.S.: Table detection from document image using vertical arrangement of text blocks. International Journal of Contents (2015)
28. Yang, J., Lu, J., Batra, D., Parikh, D.: A faster Pytorch implementation of Faster R-CNN (2017)
29. Zhong, X., ShafieiBavani, E., Yepes, A.J.: Image-based table recognition: data, model, and evaluation. arXiv (2019)
30. Zhong, X., Tang, J., Yepes, A.J.: PubLayNet: largest dataset ever for document layout analysis. In: ICDAR (2019)