ICDAR 2019 Competition on Harvesting Raw Tables from Infographics (CHART-Infographics)

Kenny Davila, Bhargava Urala Kota, Srirangaraj Setlur, Venu Govindaraju
CUBS, Dept. of Computer Science
University at Buffalo, SUNY, Buffalo, NY (USA)
Email: {kennydav, buralako, setlur, govind}@buffalo.edu

Christopher Tensmeyer, Sumit Shekhar, Ritwick Chaudhry
Adobe Research
San Jose, USA
Email: {tensmeye, sushekha, rchaudhr}@adobe.com

***Abstract*—This work summarizes the results of the first Competition on Harvesting Raw Tables from Infographics (ICDAR 2019 CHART-Infographics). The complex process of automatic chart recognition is divided into multiple tasks for the purpose of this competition, including Chart Image Classification (Task 1), Text Detection and Recognition (Task 2), Text Role Classification (Task 3), Axis Analysis (Task 4), Legend Analysis (Task 5), Plot Element Detection and Classification (Task 6.a), Data Extraction (Task 6.b), and End-to-End Data Extraction (Task 7). We provided a large synthetic training set and evaluated submitted systems using newly proposed metrics on both synthetic charts and manually-annotated real charts taken from scientific literature. A total of 8 groups registered for the competition out of which 5 submitted results for tasks 1-5. The results show that some tasks can be performed highly accurately on synthetic data, but all systems did not perform as well on real world charts. The data, annotation tools, and evaluation scripts have been publicly released for academic use.**

***Keywords*-Chart Recognition; Document Analysis; Text Recognition and Classification; Graphics Recognition; Performance Evaluation.**

I. Introduction

Charts are an effective data visualization tool often used to supplement textual content. They communicate information more efficiently and are common in the media, business documents, and scientific publications. Some charts are intended to convey a message in a visually pleasing manner, while others are used to simply display raw data, resulting in varying levels of design complexity. A considerable amount of attention has been devoted to automatically decompose and understand these visualizations [1, 2, 3, 4]. However, there is a lack of benchmarks which can be used to compare the effectiveness of the different methods proposed so far. In this work, we describe the first edition of the Competition on Harvesting Raw Tables from Infographics (ICDAR 2019 CHART-Infographics), which we consider a major step in providing common benchmarks and tools for the chart recognition community.

To provide a new common ground for evaluation, we have proposed the usage of a large-scale synthetic dataset generated from real world data sources for training and testing, and a smaller dataset based on real charts from PubMedCentral[1] as described in Section II. The complex goal of data extraction from charts has been split into multiple tasks as described in Section III. While classification tasks are evaluated with standard metrics, we propose new metrics for the chart-specific tasks. The participants of the competitions and brief descriptions of their systems are provided in Section IV, and the corresponding result of these systems for each task are analyzed in Section V. Finally, we provide our concluding remarks and recommendations for future work in Section VI. The data, annotation tools, and evaluation scripts have been publicly released for academic use[2].

Table I: Distribution of different chart types on the competition training and testing sets.

	Synthetic		PMC
Chart Type	Train	Test	Test
Pie	7,001	200	107
Donut	21,202	200	0
Line	41,874	1,000	2,257
Scatter	41,703	550	532
Vertical Box	20,958	48	69
Horizontal Box	21,007	52	4
Vertical Bar (Grouped)	8,560	595	1,207
Vertical Bar (Stacked)	13,401	625	
Horizontal Bar (Grouped)	8,586	650	66
Horizontal Bar (Stacked)	13,718	620	
Total	198,010	4,540	4,242

II. Data

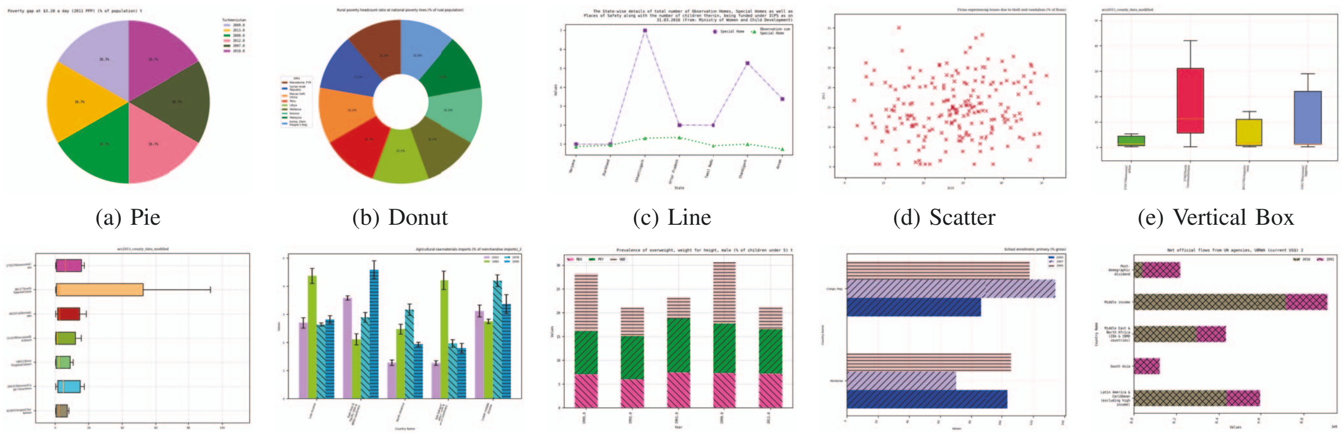
We constructed two types of datasets for this competition. The first one was a large-scale synthetic chart dataset, which allowed the participants to train and test their system using deep models. The second dataset was curated from real charts in scientific publications from PubMedCentral. Both are described in the following sub-sections.

A. *Synthetic Chart Dataset*

The Synthetic chart dataset was constructed using data tables from the following online sources: (1) World Develop-

[1]https://www.ncbi.nlm.nih.gov/pmc/
[2]https://chartinfo.github.io/

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00203



(a) Pie (b) Donut (c) Line (d) Scatter (e) Vertical Box

(f) Horizontal Box (g) Vertical Bar Grouped (h) Vertical Bar Stacked (i) Horizontal Bar Grouped (j) Horizontal Bar Stacked

Figure 1: The 10 types of charts used in the competition.

ment Indicators (World Bank)[3], (2) Gender Statistics (World Bank)[4], (3) Government of India Open Data[5], (4) Commodity Trade Statistics (UN data)[6], (5) US Census Data (for 2000 and 2010)[7], (6) Price Volume Data for Stocks and ETFs[8].

Multiple charts of different types (see Table I) were created using the Matplotlib library[9]. All tabular data was first cleansed and converted into a common format. From each table, different sets of columns and randomized rows were selected to create the charts. To emulate features of real-world charts, we introduced variations in chart component such as: (1) positioning of titles, legends, and legend entries; (2) font families and sizes; (3) style including color and/or width of lines, borders, grids, and markers; (4) bar widths and inner/outer radii of pies; and the (5) addition of optional elements such as error bars.

Afterwards, the required annotations were obtained using functions provided by Matplotlib API including tight bounding boxes for text regions, axes, legends, and data elements (e.g. bars, lines and pies). The statistics for the training and test data are provided in Table I. Samples from the dataset are shown in Fig. 1, and the dataset is publicly available at http://tc11.cvc.uab.es/datasets/CHART2019-S_1.

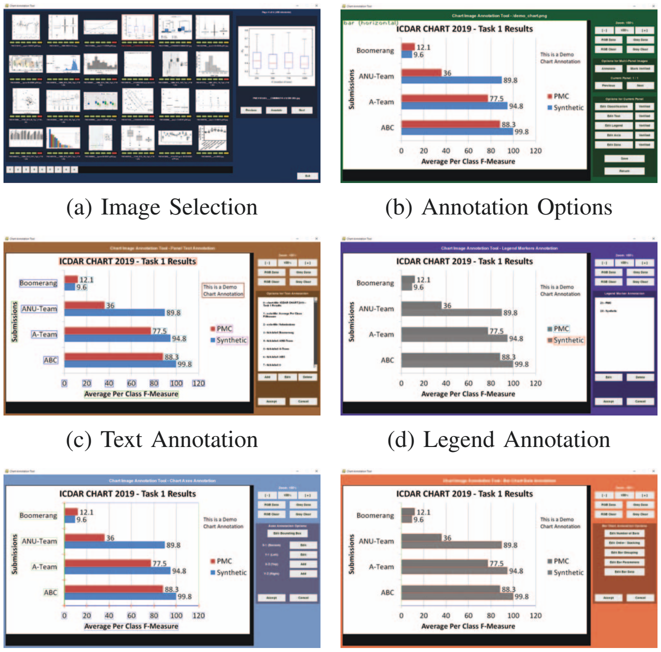
B. PubMedCentral Chart Dataset

In order to evaluate participant systems using real charts, we extracted charts from the PubMedCentral Open Access repository which contains more than 1.8 million papers. From these we extracted a small sample of charts that we have annotated in detail. This dataset is publicly available at http://tc11.cvc.uab.es/datasets/ICDAR-CHART-2019-PMC_1.

Data Sampling. We first selected journals from fields such as epidemiology, public health, pathology and genetics which were deemed to be more likely to contain charts. We only considered journals with more than 500 publications and we extracted all papers published after the year 2000. The figures from these publications were then clustered by visual similarity thus achieving a quicker separation between chart and non-chart candidates. Clusters which appeared to contain a large number of charts were selected for manual annotation of each image. We finally sampled $4,242$ single panel figures containing different chart types as described in Table I. The entire set of charts was used for evaluation of Task 1. Two disjoint sub-samples were fully annotated for evaluation of Task 2 and Tasks 3 to 5.

Data Annotation. We developed tools to annotate images based on the requirements of the different competition tasks (see Figure 2). First, the figures extracted from scientific papers were annotated with multiple labels including single vs multi panel, figure type (chart vs non-charts classes), and selected single panel chart images were further annotated by chart type and orientation (horizontal or vertical). While our tools support panel-wise annotation for multi-panel figures, we focused on single panel images for this competition. The next step was the annotation of text regions including their location, role and transcription. We used the Tesseract OCR [5] to transcribe most of the selected text regions.We manually corrected any OCR errors, and we used LaTeXstrings to represent special symbols on the text regions. After text, we annotated legends whenever these were present on the charts. Then, for axes we annotated their types (categorical vs numerical), bounding box of the first quadrant, the locations of the ticks on each axis, titles and labels of each axis, as well as any links between ticks and corresponding tick labels.

[3]www.datacatalog.worldbank.org/dataset/world-development-indicators
[4]www.datacatalog.worldbank.org/dataset/gender-statistics
[5]www.visualize.data.gov.in
[6]www.kaggle.com/unitednations/global-commodity-trade-statistics/data
[7]www.kaggle.com/muonneutrino/us-census-demographic-data/data
[8]www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs
[9]www.matplotlib.org/



(a) Image Selection (b) Annotation Options

(c) Text Annotation (d) Legend Annotation

(e) Axis Annotation (f) Data Annotation

Figure 2: Tools developed for annotation of chart images. Viewing in digital version is strongly recommended.

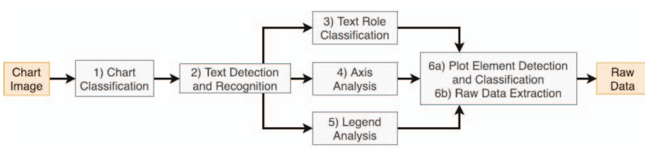


Figure 3: Pipeline of tasks for data extraction from chart images. Upstream tasks feed in their output as the input of downstream tasks.

Finally, based on the type of the chart, we developed different tools to annotate their main data elements such as bars, lines, boxes and scatter marks. Our tools are open source and are available after the competition at https://github.com/kdavila/ChartInfo_annotation_tools.

III. Tasks and Metrics

The overall task of data extraction from the chart image can be broken into a pipeline of smaller tasks (Figure 3), which we describe here along with the metrics used to evaluate the submitted systems. For each task, the ideal outputs of the previous tasks were provided to the submitted systems so that we can analyze each task independently of errors made in previous tasks.

The originally proposed tasks include 1) chart classification, 2) text detection and recognition, 3) text role classification, 4) axis analysis, 5) legend analysis, 6) Data Extraction, and 7) End-to-end Data Extraction. Of these 7 tasks, we received submissions only for Tasks 1-5, so we omit lengthy descriptions of Tasks 6 and 7. The data annotations for these tasks are distributed with the rest of the dataset.

A. Task 1. Chart Classification

In this task, chart images are classified into *horizontal bar*, *vertical bar*, *horizontal box*, *vertical box*, *line*, *scatter*, *pie*, and *donut*. Note that pie and donut charts are only used in Task 1. The metric is the average per-class F-measure. We compute the precision and recall for each class, take the harmonic mean to produce class F-measures, and then average these values to produce the final score.

B. Task 2. Text Detection and Recognition

We asked participants to perform text detection and recognition at the logical element level, e.g. multi-line and multi-word titles or axis tick labels were detected as a single element. Participants systems were provided with the chart image and the correct chart class.

Detected bounding boxes were compared to ground truth bounding boxes and considered a match if the intersection over union (IOU) exceeded a threshold (0.5). Many-to-one and one-to-many matches were resolved by choosing the pairs with highest overlaps and counting the other boxes as false negatives or false positives. Matched pairs are scored by IOU for detection as well as $\max(1 - \text{NCER}, 0)$ for recognition, where, NCER is the normalized character error rate measured as the edit distance between ground truth string and predicted string normalized with respect to ground truth string length.

Per image, the detection scores are averaged by maximum of number of ground truth boxes or number of predicted boxes, and recognition scores are averaged by number of ground truth boxes. Finally, the harmonic mean of the detection and recognition scores averaged across the entire test data set is presented as the evaluation metric for Task 2.

C. Task 3. Text Role Classification

For this task, text bounding boxes and transcripts were provided as input for all logical elements, and the participant systems were expected to classify each text element according to its semantic role in the chart. The classes are: *chart title*, *axis title*, *tick label*, *legend title*, and *legend label*. Like Task 1, we use the average per-class F-measure.

D. Task 4. Axis Analysis

For axis analysis, systems associate tick labels with pixel coordinates. Accurately understanding the axes allows chart data point geometries (bar heights, scatter points, boxes, etc.) to be translated from pixel coordinates to semantic values. The input for Task 4 is the same as Task 3 (note that which text elements are tick labels are *not* given), and the output is, for both the x-axis and y-axis, a list of text elements (tick labels), each paired with an (x, y) point representing where the tick occurs in the image.

The metric for this task is a weighted F-measure, where predicted tick marks receive a *partial true positive score* between 0 and 1 based on location accuracy. If a non-tick text element is predicted as a tick, then it receives a score of 0. We define two distance thresholds, $a = 1.0\%$ and $b = 2.0\%$ of the image diagonal, and score predictions that are a distance d from the GT tick location by

$$s(d, a, b) = \left\{ \begin{array}{ll} 1, & \text{for } d \leq a \\ \frac{b-d}{b-a}, & \text{for } a \leq d \leq b \\ 0, & \text{for } d \geq b \end{array} \right\} \quad (1)$$

Recall is computed as the sum of the scores divided by the number of GT ticks, and precision is the sum of scores divided by the number of predicted ticks.

E. Task 5. Legend Analysis

For legend analysis, we wish to pair the textual labels in the legend with the graphical marker in the legend. We take a similar formulation as Task 4, and require that systems produce a list of text elements that are paired with bounding boxes, where the bounding box (BB) surrounds the legend marker. As in Task 4, the input includes all text elements, not just those related to the legend. We score this tasks using a weighted F-measure, similar to Task 4, where true positive predictions must have the text element, but the partial true positive score is determined by BB IoU.

F. Task 6. Data Extraction

Given that charts are intended to visually convey numerical data, one main goal of chart understanding is to extract the original data (e.g. a CSV) used to create the chart. Though we received no submissions for this task, we provide a brief descriptions. We broke this task into 2 sub-tasks: 6a) Plot Element Detection and Classification and 6b) Data Extraction. Both sub-tasks received as input the ideal outputs of Tasks 1-5. For Task 6a, the goal was to segment the image into the atomic elements that represent data such as bars, points, lines, and boxes. For Task 6b, the goal was to produce named sequences of (x, y) pairs that were used to create bar, scatter and line plots, and the summary statistics for box plots.

G. Task 7. End-to-End Data Extraction

We had originally designed a task equivalent to task 6b, except that the only input was the chart image. However, received no submissions for this task. This corresponds to the real-world scenario where no GT data is given apriori.

IV. Participants

1) ABC image processing algorithm: Chen Chen, Cui Chao, Jin Song, Yang ManYe, Guo Meng from ABC Fintech. *Task 1:* ResNet-101 [6] based neural network with multi-label classification was adapted for charts. *Task 2:* Connected component analysis and Faster-RCNN [7] network with suitably modified anchor boxes was used for detection. Attention-based modules were used to handle text in multiple orientations and complex layouts for recognition. *Task 3:* Gradient boosting decision tree trained on 20 features composed from aspect ratio, position, number of horizontally/vertically aligned textboxes, whether text is numeric, direction of text, relative position between text and axis/legend was used. *Task 4:* Axis detection was done using color and line segment detection. Ticks were detected using gradient changes along axes. *Task 5:* Multi-instance segmentation model based on ResNet [6] and FPN [8] was used to segment legend markers. Rules based on aspect ratio, area and length of bounding box sides were used to eliminate markers which were matched to text boxes based on proximity. Additional custom synthetic data was generated to augment the provided data.

2) ANU-Team: Hanif Rasyidi from Australian National University, Research School of Electrical, Energy and Materials Engineering. *Task 1:* Neural network with ResNet-50 [6] architecture, pretrained on ImageNet [9], with a dense layer regularized using Dropout was used.

3) A-team: Joao Pinheiro from Universidad Catolica San Pablo, Peru and Jorge Poco from Fundação Getulio Vargas, Brazil. *Task 1:* Neural network with VGGNet architecture [10], pretrained on ImageNet [9] was used. *Task 2:* PixelLink text detection system [11], pretrained on ICDAR 2015 scene text dataset [12], was finetuned with the training data provided. For text recognition, Tesseract OCR [5] was used. *Task 3:* SVM trained on geometric features extracted from text bounding boxes was used [13]. *Task 4:* Using tick label text boxes, tick points were located based on proximity of text to a 'skeleton' constructed from the input image. *Task 5:* The original image was binarized and connected components are extracted and filtered by size. A component was matched to a legend label based on proximity as the legend marker. For PMC data, additional training data [13, 14] was used along with a fraction of synthetic data for Tasks 1-2.

4) Boomerang: Rima Hazra, Atharva Vyas from Indian Institute of Technology Kharagpur, India. *Task 1:* A neural network with 5 convolutional layers with ReLU activation (2 $\times$ 32 depth with stride 5, 2 $\times$ 64 depth with stride 5 and 1 $\times$ 128 depth with stride 5) followed by a dense layer (1024 neurons) with Dropout regularization was used.

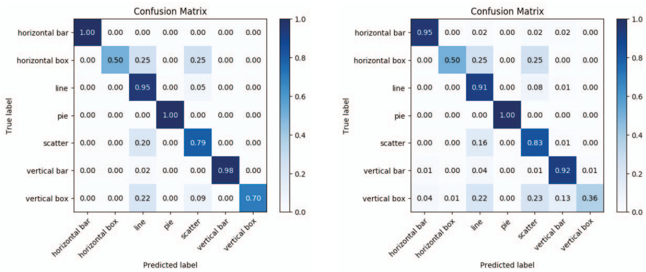
5) IITB-Team: Utkarsh Gupta, Soumen Chakraborthi from Indian Institute of Technology Bombay, India. *Task 3:* Multi-class classification was performed using an SVM with one vs one classification strategy for text role classification. From each input bounding box, 14 features are extracted derived from aspect ratio, coordinates, position with respect to center of image, overlaps with other boxes and position relative to a container bounding box.

V. Competition Results

In the following sub-sections we analyze the performance of the submitted systems on the tasks.

Table II: Results for Task 1 (average F-measure)

Team	Synthetic	PMC
ABC	**99.81**	**88.29**
A-Team	94.82	77.52
ANU-Team	89.78	35.96
Boomerang	9.59	12.06



(a) ABC (b) A-team

Figure 4: Task 1 Confusion Matrices for the top 2 performing systems on PMC.

A. Chart Classification Results

We received the most submissions for Task 1, with 4 participants submitting results for both the Synthetic and PMC dataset. In general, accurate results were obtained on the synthetic data, with nearly perfect accuracy obtained by the ResNet classifier employed by ABC. The main source of confusion for ABC and A-team were classifying bar charts as either stacked or grouped.

For the PMC dataset, ABC still performed best, but all teams had much higher error rates compared to the Synthetic dataset. The confusion matrices (Fig. 4) for the 2 best performing methods look similar. In particular, both horizontal and vertical box plots were often misclassified as line or scatter plots, and line plots were difficult to distinguish from scatter plots. This may be because some scatter plots in PMC do have lines (e.g. a best fitting line), but the lines themselves do not represent the raw data. For ABC, horizontal bar, vertical bar, and pie charts were classified with near perfect accuracy.

It seems the key to obtain high performance on PMC was to use additional data beyond the Synthetic training data. For ABC, this additional data was created by applying many data augmentation transformations to the synthetic images, which helped the classifier be more robust. For A-team, the model was pre-trained on ImageNet [9] and fine-tuned on some synthetic data and 2 external chart datasets [13, 14]. ANU-Team only pre-trained on ImageNet and fine-tuned on the provided Synthetic dataset, so the resulting classifier was less robust to the real-world charts.

B. Text Detection and Recognition Results

Only two teams submitted results for Task 2, text detection and recognition, as it is presented in Table III. Both teams

Table III: Results for Task 2

	Synthetic		
Team	IOU	OCR	F-measure
ABC	69.92	94.97	80.54
A-Team	70.96	78.97	74.75
	PMC		
A-Team	48.48	58.81	53.15

submitted results for the Synthetic test set and only A-team submitted results for the PMC test set. We notice that both approaches achieve similar performance in terms of text detection as shown by their very similar scores in IOU. However, in terms of text recognition we can see that team ABC achieves a considerably much higher score which leads to an overall higher F-measure score. While only one team submitted results for this task on the PMC dataset, we can see that their numbers are considerably lower for this dataset. We found this dataset in general to be more challenging for text detection since images are not always available in an appropriate resolution, and the text regions are sometimes hard to read even for human annotators. Many text regions in this dataset also contain a considerable number of special symbols such as Greek letters, superscripts, and subscripts. It is worth noting that while A-team used the same OCR engine (Tesseract OCR) that was employed to speed up the annotation process, we found that this engine failed to recognize a large number of strings that had to be manually typed in during the ground truth generation process.

C. Text Role Classification Results

Three teams submitted results for Task 3 as shown in Table IV. It is important to notice that the class *legend title* was present in only a few images in the PMC test set and was not present in the Synthetic dataset. Therefore, we decided to omit it from the Averaged per-class F-measure. We also omitted two more roles (*Value Label* and *Other*) that were present on the PMC dataset but not present in the Synthetic data set. The final numbers provided are the average F-measure for *Chart Title*, *Axis Title*, *Legend Label* and *Tick Label*. Both A-team and ABC team produced nearly perfect scores for the Synthetic dataset, but we can see that the score for the PMC submissions left room for considerable improvement. Common confusion in this dataset were related to text regions from different classes being misclassified as legend labels due to their similarity in layout organization.

D. Axis Analysis Results

Only two teams submitted results for Task 4 as shown in Table IV. A-Team achieved a near perfect weighted f-measure score on the Synthetic dataset, while ABC team also achieved a very high score. All ticks in the Synthetic dataset followed a single location pattern were each tick was directly associated with a label. However, for the PMC dataset there is a much larger variety in terms of ticks styles

Table IV: Results for Tasks 3, 4 and 5

	Synthetic		
	Task 3	Task 4	Task 5
Team	Average F-measure	Weighted F-measure	Weighted F-measure
ABC	100.00	96.49	78.14
A-Team	99.95	99.76	87.13
IITB-Team	60.25	-	-
	PMC		
A-Team	84.38	77.33	-
IITB-Team	35.58	-	-

and locations. For example, many charts include minor ticks that are only used as a visual aid but are not linked to any tick label. In this category, we also find ticks whose main purpose is to separate chart regions and are usually located in-between tick labels. Such ticks were explicitly annotated on the PMC dataset, but since these were not present on the Synthetic dataset, the submission by A-team did not consider them. We decided to modify our ground truth to produce annotations that followed the only pattern available on the synthetic data, and finally we evaluated A-team submission based on this modified version of the ground truth.

E. Legend Analysis Results

We received 2 submissions (ABC and A-team) for Task 5 on the Synthetic dataset and the results are shown in Table IV. In this case, A-team (87.13%) outperformed ABC (78.14%).

Given that legend analysis involved identifying which text elements are legend labels and finding the corresponding marker bounding boxes (BBs), there are 2 sources of errors. However, as evidenced by the near perfect performance of both methods on Task 3, neither system made any errors in identifying legend labels. Therefore, the difference in performance is entirely due to localization of the marker bounding boxes. While ABC had a higher percentage of predicted BBs with IoU > 0.95 (13.7% vs 1.3%), they also had a much higher percentage of BBs that had 0.0 IoU (17.9% vs 3.5%). In general, ABC had a wide variety of BB quality, while most of the A-team BBs were good. For ABC, 44.7% of BBs had IoU > 0.80, while 78.7% of A-team's BBs met this criteria.

F. Tasks 6 and 7 Results

Unfortunately, none of the participant teams were able to produce results for these tasks for any of the test sets. This may indicate that the overall chart parsing task is challenging.

VI. Conclusion

In this competition, we have provided a new benchmark for the chart recognition community. From the different results on our tasks and the usage of both real and synthetic data we have learned that current systems are able to handle synthetic data from a single source very well, but real charts are much more challenging. In particular, we consider that two teams (ABC and A-team) achieved a tie for the Synthetic dataset, with each performing best on 2 tasks, and performing roughly equally on Task 3. A-team is the winner for the PMC dataset, outperforming one other system on Task 3 and being the only submission for Tasks 2 and 4.

In the future, we would like to consider different options to increase the variety of the synthetic data to reflect more of the challenges found in real charts. We also plan to construct larger manually annotated datasets of real charts. The proposed overall Tasks 6 and 7 were challenging as shown by the lack of participation on these tasks. However, we hope that the release of fully annotated data and evaluation scripts for these will help to produce systems which can handle these tasks in the future.

Acknowledgment

This material was partially supported by the National Science Foundation under Grant No. 1640867 (OAC/DMR). We would also like to acknowledge Utkarsh Gupta, Prann Bansal and Ajay Shridhar Joshi, Adobe Research Summer 2018 Interns and Pranav Maneriker, ex-Adobe Research Fellow for their contribution to synthetic data generation.

References

[1] H. C. Purchase, "Twelve years of diagrams research," *Journal of Visual Languages & Computing*, vol. 25, no. 2, pp. 57–75, 2014.

[2] Y. Liu, X. Lu, Y. Qin, Z. Tang, and J. Xu, "Review of chart recognition in document images." in *Visualization and Data Analysis*, 2013, p. 865410.

[3] H. Mei, Y. Ma, Y. Wei, and W. Chen, "The design space of construction tools for information visualization: A survey," *Journal of Visual Languages & Computing*, 2017.

[4] F. Böschen and A. Scherp, "A comparison of approaches for automated text extraction from scholarly figures," in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 15–27.

[5] R. Smith, "An overview of the Tesseract OCR engine," in *International Conference on Document Analysis and Recognition*, vol. 2. IEEE, 2007, pp. 629–633.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[8] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in *CVPR*, vol. 1, no. 2, 2017, p. 4.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *AAAI Conference on Artificial Intelligence*, 2018.

[12] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "ICDAR 2015 competition on robust reading," in *International Conference on Document Analysis and Recognition*. IEEE, 2015, pp. 1156–1160.

[13] J. Poco and J. Heer, "Reverse-Engineering visualizations: Recovering visual encodings from chart images," in *Computer Graphics Forum*, vol. 36, no. 3. Wiley Online Library, 2017, pp. 353–363.

[14] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer, "Revision: Automated classification, analysis and redesign of chart images," in *Symposium on User Interface Software and Technology*. ACM, 2011, pp. 393–402.