2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

ICPR2018 Contest on Robust Reading for Multi-Type Web Images

Mengchao He*[1], Yuliang Liu*[2], Zhibo Yang*[1], Sheng Zhang[2], Canjie Luo[2],
Feiyu Gao[1], Qi Zheng[1], Yongpan Wang[1], Xin Zhang[2] and Lianwen Jin[2]
[1]Alibaba Group
[2]South China University of Technology
Emails: zhibo.yzb@alibaba-inc.com; mengchao.hmc@alibaba-inc.com; eelwjin@scut.edu.cn

***Abstract*—Electronic commerce has infiltrated every aspect of our daily lives, which offers great convenience for shopping, advertising, etc. Text in the web images is responsible to convey essential information for consumers. Algorithms that read text in these web images can facilitate applications of various types, such as goods surveillance, products classification, and intelligent retrieval or recommendation. Despite of various existing text reading tasks, this contest introduces a novel large-scale dataset named MTWI that contains 20,000 images, which is the first dataset that is mainly constructed by Chinese and English web text. Three tasks (web text recognition, web text detection, and end-to-end web text detection and recognition) were set up for encouraging more research on the web text reading problem. The contest was held from February 2, 2018 to May 26, 2018 with 289 valid submissions from 4,282 registered teams. Throughout this report, we describe the details of this new dataset, the purposes and definitions of the tasks, the evaluation protocols, and the summaries of the results.**

I. Introduction

Electronic commerce offers great convenience for shopping or advertising, which is now becoming an indispensable part of our life. Driven by the increasing demand of consumers, more and more products are available for online-shopping, thousands of web images have been generated simultaneously. Text in these web images is one of the most essential forms to convey valuable information. The web text ability can facilitate applications of various kinds, such as goods surveillance, products classification, content filtering and intelligent retrieval or recommendation.

Retrospectively, the text reading problem is one of the main concerns. In the past few years, lots of successful text detection and recognition methods and systems come out, which were mainly benefited from the standard benchmarks. Before 2011, most publicly available datasets contained little or even no web text, and the web text reading had been less studied in the community. It was not until ICDAR2011 contest "Reading text in born-digital images (web and email)" [8] had firstly introduced web text reading problem, which illuminated the large potential value of reading web text. Since then, it seemed not other relevant datasets were built for web text. Many large and challenging datasets are constructed to target many kinds of scene text, such as ICDAR2013 "Focused Scene Text" [9], ICDAR2015 "Incidental Scene Text" [7], RCTW-17 [18] the largest Chinese scene text datasets, and MLT [15] the largest multi-lingual scene text datasets.

Therefore, regarding the value of reading web text, we proposed this contest with a magnitude larger dataset than [8]. There are also many other improvements of our dataset compared to [8]: 1) most images of our dataset have relatively high resolution with more text instances.; 2) images of our dataset cover more types of web text, including fancy advertising-fonts, multi-lingual texts from all over the world etc. 3) the dataset has many multi-oriented text, tightly-stacking text and complex-shaped text, such as distorted text and curved text. All these challenges and characteristics make the proposed contest a novel and unique problem.

The proposed dataset includes 20,000 carefully-selected images with a large-scale web text. The images in this dataset were carefully annotated with quadrangles and text transcriptions. Some unrecognizable small text regions were also been annotated but the transcriptions were marked as "###", which indicates "don't care". The contest is divided into three tasks: the web text recognition, web text detection, and end-to-end web text detection and recognition.

The contest has been held from February to the end of May, 2018. We have received extensive attentions from the related research community. In total, 1370, 1424, and 1488 teams have registered for task 1, 2, and 3, respectively. Among them, there are 70, 193, and 26 valid results for each task. In the following, we present details of the proposed dataset, definitions of tasks and evaluation protocols, and summarization of submitted methods.

II. Dataset and Annotations

The dataset, i.e., MTWI, has 20,000 web images (half as the training set and half as the testing set), which is selected based on the business properties and market division of the Taobao[1] front-end. We have defined 17 categories from current business types of Taobao. According to the proportion of each category, we first collected 200,000 data as our initial samples, and the number of each category is listed in Figure 1. Subsequently, we manually selected 20,000 samples for the contest including the following scenes, typography, and designs:

*Authors contributed equally as first author.

[1]https://www.taobao.com/

978-1-5386-3787-6/18/$31.00 ©2018 IEEE

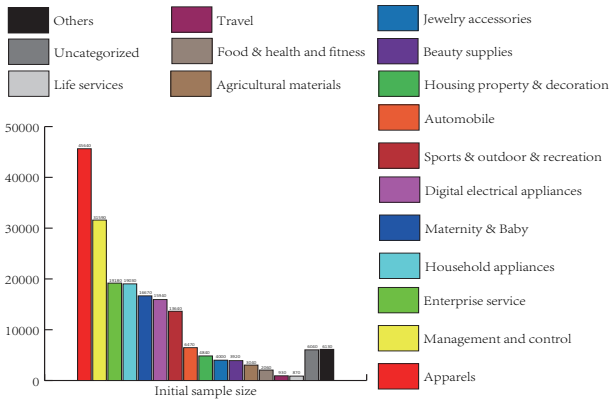


Fig. 1. Number of initial samples in each category.

- **Scenes**. Digital-born text, text on the surface of the objects, printed text in the cover, text in electronic device, text in web pages, scene text, etc.
- **Typography**. Single, horizontal, vertical, oblique, curvilinear, complex text (mixed layout with different font sizes, different kinds of fonts, different word spacing), occlusion, etc.
- **Designs**. Common fonts, art design characters, watermark characters, stroke characters, hollow characters, stencils, logo characters, etc.

The overall selecting principle is that samples have not or less been covered by previous datasets. Although the images are mainly from commercial web sites, the types of web images are still various, including born-digital images, web page illustrations, scene text (for live demonstration of the goods), etc.

After collecting the original image data, it took us approximately two months to label the entire dataset. In the bounding box annotation procedure, we adopt the following rules: 1) each text, whether it is recognizable or not, should be tightly annotated by a quadrangle; 2) for each bounding box, if the word space exceeds one character, then it is divided into two bounding boxes. For the recognition annotation (literal transcription), some basic rules are also adopted: (1) case sensitive letters and Chinese and English symbols are distinguished; (2) traditional Chinese characters are labeled as traditional Chinese characters; (3) Japanese, Korean and other non-Chinese characters only need to be localized without transcription; (4) unrecognizable text bounding box does not need to be annotated with transcription.

The texts were carefully annotated and the string length distribution is shown in Figure 2.

III. Challenge Tasks and Evaluating Protocol

We set up three contest tasks: cropped web text recognition task, web text localization and end-to-end web text detection and recognition.

A. Task 1 - Web Text Recognition

Text recognition is a conventional contest task, which aims at recognizing the content of each text instances. This task is challenging mainly due to the large number of classes of Chinese category and various of fancy web text fonts. Following the previous RCTW-17 [18], the evaluation protocol adopts a traditional evaluation metric: the Normalized Edit Distance (NED). The score is calculated by 1-NED. The competitor is allowed to use extra training data, including synthetic text.

B. Task 2 - Web Text localization

Text detection is commonly considered to be a prerequisite task for text recognition. This task aims to evaluate the text localization performance. Because of about 10% inconsistent annotation granularity, the evaluation protocol in ICDAR 2013 [9], [19] is adopted, which uses one-to-many and many-to-one metrics to be compatible with annotation granularity. This protocol is more suitable to our dataset since the word-level or line-level text bounding box can both be recognized as well. In order to restrict over-segmentation and encourage fine-grained detection, we use a penalty function following the same as [19], which corresponds to the fragmentation index suggested by Mariano et al. [14], as shown below:

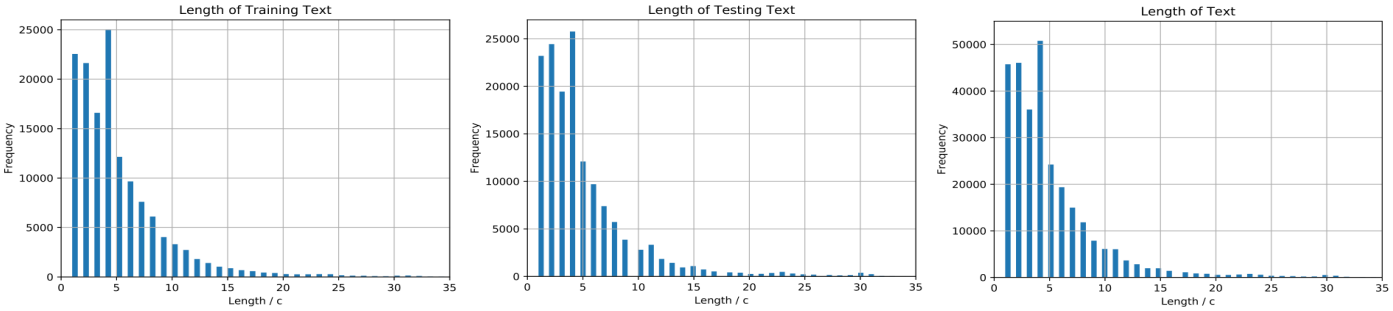
$$f_sc(k) = \frac{1}{1 + ln(k)}, \tag{1}$$

where, k is the number of matched detected boxes (for many-to-one, it represents the number of the matched ground truth (GT) boxes). More details can be found on [19]. In addition, because a little information loss could cause mistaken recognition, thresholds of item recall and item precision [19] are both set to 0.7 in this contest. The primary metric for this task is the harmonic average (Hmean) value of precision and recall.

C. Task 3 - End-to-End Web Text Detection and Recognition

This task requires that the method can end-to-end detect and recognize text. The submitted result format is the coordinates of detecting bounding box together with recognition results, and the recognized text is required to be encoded as UTF-8 strings. The evaluation protocol mainly follows the concepts of RCTW-17 [18].

The normalized edit distances (NED) between recognized text and ground truth (GT) text are used to evaluate the performance of each method. According to the matched rules in recall and precision, the corresponding NEDs are summed and divided by the number of test instances, which is called NED recall and NED precision, respectively. Based on NED recall and NED precision, the harmonic average of them can be calculated, which serves as a primary metric. The evaluation process includes two steps: 1) all detection results are matched to either a) a GT bounding box that the recall (overlapping area divides the GT area) and precision (overlapping area divides the detection area) are both higher than a threshold 0.5, or b) “None” if none of the GT bounding box is matched with the detection result. If multiple detected bounding boxes are matched to the same GT bounding box, then the top three with the maximum sum of precision and recall will



(a) Training set. (b) Testing set. (c) Total set.

Fig. 2. String length distribution of the proposed dataset.

be preserved and the rest are matched to "None". Because multiple recognition results for the same GT bounding box may be beneficial to maintain the integrity of the whole text. 2) Based on the matched matrix, all the NEDs between matching pairs can be calculated. If a detection is matched to "None", then the NED is calculated between the recognized text and an empty string. The NEDs are summed and divided by the number of test instances. This metric evaluates detection and recognition performance simultaneously. Note that detection results matched to text instances with "don't care" flags increase zero NED, which will be avoided in the same manner as [7] did.

IV. Organization

The website and registration for the contest was opened on February 10, 2018, and the training and testing data were available on March 15, 2018 and May 14, 2018, respectively. To ensure all the teams could submit valid results, we provide three times evaluation and leaderboard update at 10:00 UTC+8 on May 17, May 21 and May 23, 2018, respectively. Subsequently, the report submission is opened for top 10 teams until the deadline: May 26, 2018. Final results announcement is on June 1, 2018, and the awards ceremony is held on ICPR 2018, August 20-24, 2018.

We received 1370, 1424, and 1488 registrations where 70, 193, and 26 are valid results for task 1, 2, and 3, respectively. Most of the teams are from Chinese universities, research institutes, and technology companies. Some registrations come from universities in United States, United Kingdom, Japan and Singapore. Teams must submit their results based on the specific format. Similar to previous text robust reading contests, the detection task attracted the most attentions. Contrastively, task 1 and task 3 have much less number of valid results, which is mainly because of the complexity of the Chinese text recognition task.

It is worth mentioning that a monetary award is offered for the top three methods in our contest: for task 1 and 2, the champion, the 2nd place and 3rd place are rewarded with 15k, 10k and 5k RMB, respectively; for task 3, the rewards are 20k, 12k and 8k RMB for top three methods, respectively.

V. Submission and Results

The contest website[2] has provided sufficient information to guide and inform competitors. We evaluate all submitted results at the end of each evaluation step. The top ten results of each task are listed as Table I, II, and III, respectively. Note that according to the contest rule, any team who didn't submit their technical report would be deemed as abandon. We include the team name, affiliations, and primary metric value in the tables. As the training data is one of essential parts for a successful deep learning network, we also show how the exterior training data was used. The exterior training data don't affect the final results but may be a useful reference for researches. Methods are ranked by the primary metric of each task, and some of the submitted results are visualized in Figure 3, 4, and 5.

Because the company may have more computing sources and data, we will give brief introduction for the top 3 methods of companies and universities, respectively.

A. Top 3 submissions for Task 1 - Company

1. "nelslip(iflytek&ustc)" (IFLYTEK, Team Members: Jianshu Zhang, Mingjun Chen, Jiajia Wu, Jinshui Hu, Jun Du, Yixing Zhu, Lirong Dai and Wenchao Wang). In general, the team adopts the attention based encoder-decoder model for this challenge. The team uses DenseNet [5] as encoder [21] and employs the Radical Analysis Network (RAN) [22] to handle the large amount of Chinese characters recognition. The recognition performance is enhanced by appending an end-of-word (eow) flag after each Chinese character Ideographic Description Sequences (IDS) caption. Data augmentation is used to improve recognition performance.

2. "Samsung R&D China, Beijing" (Samsung R&D China, Beijing, Team Members: Hao Guo, Haiyang Guo, Pingjun Li, Zhenbo Luo, Yingying Jiang and Xiaobing Wang). The method mainly refers to [16]. The convolutional layers of the network is replaced by the first four block of VGG, and the pooling after conv3_3 only focus on the height dimension of

[2] https://tianchi.aliyun.com/markets/tianchi/icpr_mtwi_2018_challenge

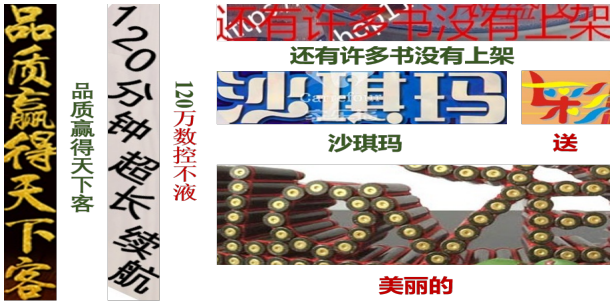


Fig. 3. Visualization of the recognition results. Green: correct recognition results. Red: wrong recognition results.

Fig. 4. Visualization of the detection results.

the input. The recurrent layers is composed of 3 Bi-LSTM, and baidu_ctc[3] is used to calculate loss.

3. "NTAI" (NetEase, Team Members: Binbin Xu and Li Lin). The method combines ResNet, bi-directional LSTM, and CTC for sequential text recognition, which is similar to the CRNN pipeline [16]. First, the team members pretrain the model on the synthesized text dataset. Then, they finetune their model only on the contest training set without any other dataset.

B. Top 3 submissions for Task 1 - Academic Institute

1. "—Baseline—" (Hebei University of Science & Technology, Team Members: Hongfei Sun, Ming Sun, Hongping Ren, Minjie Chen, Bowen Yang and Weikang Zhao). The text recognition framework (CRNN+attention) is inspired by referring to [22]. Using CNN to process images and extract features, and the output features is attention-based weighted as the input of LSTM.

2. "mclab_rec" (Huazhong University of Science & Technology, Team Members: Mingkun Yang, Hui Zhang, Zhisheng Zou, Qingquan Xu and Xiang Bai). The proposed method of recognition is based on Attention Models [17]. BLSTM is used for decoding, and multi-scale model integrity is used to further improve the results.

3. "hitsz_icrc" (Harbin Institute of Technology (Shenzhen), Team Members: Xiangping Wu, Jinghan You, Weilin Zhang, Yulun Xiao and Qingcai Chen). They use ResNet-101 [3] and BLSTM for sequential text line recognition.

C. Top 3 submissions for Task 2 - Company

1. "nelslip(iflytek&ustc)" (IFLYTEK, Team Members: Jianshu Zhang, Mingjun Chen, Yixing Zhu, Jiajia Wu, Jinshui Hu, Jun Du and Lirong Dai). The method is based on FPN [10], and the team augmented the network as PANet [11] does. SLPR [23] and two-steps Cascade R-CNN [1] are used to fit the outline of the text line and generate the quadrilateral results.

2. "Samsung R&D China, Beijing" (Samsung R&D China, Beijing, Team Members: Yi Yu, Yingying Jiang, Xiangyu Zhu, Hao Guo, Zhenbo Luo and Xiaobing Wang). The method mainly refers to [6]. The improvement points are: a) Replacing the basic feature extraction module with ResNeXt-101 [20]; b) adding FPN [10] enhancement feature extraction; c) Replacing the RoIPooling with RoIAlign. The method also uses multi-scale training and single-scale testing strategy.

3. "UC" (sensetime, Team Members: Xuebo Liu and Yingxu Wang). The team uses FOTS [12] for text detection. The backbone network is ResNet-50. Only the official training dataset is used for training.

D. Top 3 submissions for Task 2 - Academic Institute

1. "NJU_ImagineLab_PSENet" (Nanjing University, Team Members: Wenhai Wang, Cheng Wang, Xiaoge Song, Wenbo Hou and Fei Han). The method is based on their recently submitted method: segmentation network "Progressive Scale Expansion Network (PSENet)".

2. "mclab_rec" (Huazhong University of Science & Technology, Team Members: Pengyuan Lv, Minghui Liao, Mingtao Fu, Minghan He and Xiang Bai). It is a Mask R-CNN [2] based algorithm. The backbone of the model is ResNet-50 [3].

3. "Dream Team" (Nanjing University, Team Members: Bin Zhao and Shuyang Jin). It is a Mask R-CNN [2] based algorithm. The backbone of the model is ResNet-101 [3].

E. Top 3 submissions for Task 3 - Company

1. "nelslip(iflytek&ustc)" (IFLYTEK, Team Members: Jianshu Zhang, Mingjun Chen, Jiajia Wu, Jinshui Hu, Jun Du, Yixing Zhu and Lirong Dai). A combination of detection and recognition introduced in task 1 and task 2.

2. "Samsung R&D China, Beijing" (Samsung R&D China, Beijing, Team Members: Yi Yu, Yingying Jiang, Pingjun Li, Haiyang Guo, Hao Guo, Zhenbo Luo, Xiaobing Wang and Xiangyu Zhu). A combination of detection and recognition introduced in task 1 and task 2.

3. "UC" (sensetime, Team Members: Xuebo Liu, Yingxu Wang and Yichao Wu). The team combines algorithms used in Task1 and Task2 for Task3. First they detect text regions, and then crop text patch from images and using a CRNN structure for text recognition.

F. Top 3 submissions for Task 3 - Academic Institute

1. "PALQQ" (Chinese Academy of Sciences, Team Members: Wenhao He, Wei Feng and Ruiqi Wang). Detection

[3]https://github.com/baidu-research/warp-ctc

TABLE I

RESULTS SUMMARY FOR THE TOP-10 SUBMISSIONS OF TASK 1. CYAN REPRESENTS TECH COMPANY AND PINK REPRESENTS UNIVERSITY OR ACADEMIC INSTITUTE. NED: NORMALIZE EDIT DISTANCE. PM: PERFECT MATCHING, MEANING THE RECOGNITION RESULTS MUST BE TOTALLY THE SAME AS THE GT CHARACTERS OF EACH ANNOTATED BOX. S: SYNTHETIC DATA. PA: PRIVATE DATA. PB: PUBLIC AVAILABLE DATA.

Team Name	Affiliation	Score	NED	PM	Extra Data
nelslip(iflytek&ustc)	IFLYTEK	0.858	0.142	0.722	PA (250k)
SRC-B-MachineLearningLab	Samsung R&D China, Beijing	0.857	0.143	0.714	S (14m) + PB (170k) + PA (260k)
NTAI	NetEase	0.826	0.174	0.665	S (1m) + (PB + PA) (100k)
UC	sensetime	0.825	0.175	0.666	S (Unknown)
—Baseline—	Hebei University of Science & Technology	0.810	0.190	0.652	S (6m)
mclab_rec	Huazhong University of Science & Technology	0.806	0.194	0.664	S (16m)
hitsz_icrc	Harbin Institute of Technology (Shenzhen)	0.802	0.198	0.636	S (467k)
FDU____	Fudan University	0.790	0.210	0.604	S (Unknown) + PB (10k)
GUOCICI	Beijing Institute of Technology	0.752	0.248	0.561	S (80k) + PB (8k)
CCFlab	Shanghai Jiao Tong University	0.750	0.250	0.550	PB (30k)

TABLE II

RESULTS SUMMARY FOR THE TOP-10 SUBMISSIONS OF TASK 2. CYAN REPRESENTS TECH COMPANY AND PINK REPRESENTS UNIVERSITY OR ACADEMIC INSTITUTE. S: SYNTHETIC DATA. PA: PRIVATE DATA. PB: PUBLIC AVAILABLE DATA.

Team Name	Affiliation	Hmean	Precision	Recall	Extra Data
nelslip(iflytek&ustc)	IFLYTEK	0.796	0.813	0.779	-
SRC-B-MachineLearningLab	Samsung R&D China, Beijing	0.766	0.813	0.723	-
UC	sensetime	0.755	0.788	0.725	Unknown
NTAI	NetEase	0.752	0.799	0.711	-
NJU_ImagineLab_PSENet	Nanjing University	0.752	0.785	0.721	PB (10k)
mclab_det	Huazhong University of Science & Technology	0.734	0.788	0.687	S (800k) + PB (1k)
Dream Team	Nanjing University	0.732	0.749	0.716	-
XiaoZuanFeng	Dalian University of Technology	0.715	0.776	0.663	-
NaiveCrediks	Chinese Academy of Sciences	0.714	0.815	0.635	-
JiaoShenMeMingZi	Shanghai Jiao Tong University	0.708	0.781	0.648	-

adopts direct regression method [4]. A simple merge strategy is used to improve long text detection performance. Text recognition is a sliding window based method, which crops the image for single word recognition (total 7356 classes). CTC and language model are used for decoding.

2. “mclab_rec” (Huazhong University of Science & Technology, Team Members: Mingkun Yang, Hui Zhang, Zhisheng Zou, Minghan He, Pengyuan Lv, Xiang Bai, Mingtao Fu and Minghui Liao). A combination of detection and recognition introduced in task 1 and task 2.

3. “553” (Chinese Academy of Sciences, Team Members: Yuanshun Cui, Yu Song, Jie Li, Hongyu Pan and Jiyun Cui). Using FCN [13] and ResNet-50 for text detection. The detected text is coded by an attention-based recognition model (CRANN).

Fig. 5. Visualization of the end-to-end detection and recognition results.

VI. CONCLUSION

We organized the first web text reading challenge - ICPR MTWI 2018 contest. A more rigorous evaluation protocol is utilized. The contest lasted for more than 3 months, which has received extensive attention and diversified teams, indicating the text reading problem is popular in the community.

According to the results, the best performance for three

TABLE III
RESULTS SUMMARY FOR THE TOP-10 SUBMISSIONS OF TASK 3. CYAN REPRESENTS TECH COMPANY AND PINK REPRESENTS UNIVERSITY OR ACADEMIC INSTITUTE. S: SYNTHETIC DATA. PA: PRIVATE DATA. PB: PUBLIC AVAILABLE DATA.

Team Name	Affiliation	Hmean	Precision	Recall	Extra Data
nelslip(iflytek&ustc)	IFLYTEK	0.815	0.787	0.846	PA (250k)
SRC-B-MachineLearningLab	Samsung R&D China, Beijing	0.784	0.810	0.759	S (14m) + PB (170k) + PA (260k)
UC	sensetime	0.754	0.734	0.775	Unknown
PALQQ	Chinese Academy of Sciences	0.720	0.717	0.722	PB (8k)
NTAI	NetEase	0.719	0.688	0.753	S (1m) + (PB + PA) (100k)
mclab_e2e	Huazhong University of Science & Technology	0.717	0.757	0.681	S (16m)
553	Chinese Academy of Sciences	0.665	0.671	0.659	S (9.8m) + PB (40k)
origins	Shanghai Jiao Tong University	0.634	0.704	0.576	PB (9k)
LIBBLE	Nanjing University	0.633	0.666	0.603	S (Unknown)
CCFlab	Shanghai Jiao Tong University	0.627	0.571	0.695	PB (30k)

tasks are all less than 90%, indicating there are insufficiencies to be improved. Through our inspection, we summarize some reasons that may affect the results: a) For task 1, the fancy text fonts and ambiguous Chinese characters are one of the main reasons that result in false recognition. b) For task 2, complex-shaped text, such as curved and distorted text, is one of the main reasons causing unsatisfactory detection. c) The performance of End-to-End task is heavily dependent on the text detection performance. In addition, the way to handle the engagement between detection and recognition is also important for end-to-end recognition.

More details about the results can be found on the website. In the future, we will maintain and improve the dataset to further make contributions to the research community.

ACKNOWLEDGEMENT

The authors thank Dr. Cheng-Lin Liu, Dr. Xiang Bai and Dr. Fei Yin for their suggestions. This challenge is sponsored by Alibaba Group, and is supported in part by NSFC (61673182) and GD-NSF (no.2017A030312006).

REFERENCES

[1] Z. Cai and N. Vasconcelos. Cascade R-CNN: Delving into high quality object detection. *arXiv preprint arXiv:1712.00726*, 2017.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu. Deep direct regression for multi-oriented scene text detection. *Proceedings of the IEEE International Conference on Computer Vision*, pages 745–753, 2017.

[5] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, pages 4700–4708, 2017.

[6] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo. R2CNN: Rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.

[7] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, et al. ICDAR 2015 competition on robust reading. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

[8] D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy. Icdar 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email). In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1485–1490. IEEE, 2011.

[9] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. I. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazn, and L. P. D. L. Heras. ICDAR 2013 robust reading competition. In *International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, pages 2117–2125, 2017.

[11] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. *arXiv preprint arXiv:1803.01534*, 2018.

[12] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan. FOTS: Fast oriented text spotting with a unified network. *CVPR*, 2018.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[14] V. Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer. Performance evaluation of object detection algorithms. In *16th International Conference on Pattern Recognition*, volume 3, pages 965–969. IEEE, 2002.

[15] N. Nayef, F. Yin, I. Bizid, H. Choi, et al. ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *14th International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1454–1459. IEEE, 2017.

[16] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2017.

[17] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4168–4176, 2016.

[18] B. Shi, Yao, M. Liao, Y. M., X. P., L. Cui, L. S. Serge Belongie, and B. X. ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17). *arXiv preprint arXiv:1708.09585*, 2017.

[19] C. Wolf and J. M. Jolion. Object count/area graphs for the evaluation of object detection and segmentation algorithms. *International Journal of Document Analysis & Recognition*, 8(4):280–296, 2006.

[20] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 5987 5995, 2017*, pages 5987–5995. IEEE, 2017.

[21] J. Zhang, J. Du, and L. Dai. Multi-scale attention with dense encoder for handwritten mathematical expression recognition. *arXiv preprint arXiv:1801.03530*, 2018.

[22] J. Zhang, Y. Zhu, J. Du, and L. Dai. RAN: Radical analysis networks for zero-shot learning of chinese characters. *arXiv preprint arXiv:1711.01889*, 2017.

[23] Y. Zhu and J. Du. Sliding line point regression for shape robust screne text detection. *arXiv preprint arXiv:1801.09969*, 2018.