ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT

Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al.

ICDAR2017 Robust Reading Challenge on Multi-lingual Scene Text Detection and Script Identification – RRC-MLT

Nibal Nayef[⋆1], Fei Yin[†1], Imen Bizid[⋆2], Hyunsoo Choi[⊙2], Yuan Feng[†2], Dimosthenis Karatzas[•2], Zhenbo Luo[◇2], Umapada Pal[‡2], Christophe Rigaud[⋆2], Joseph Chazalon[⋆3], Wafa Khlif[⋆3], Muhammad Muzzamil Luqman[⋆3], Jean-Christophe Burie[⋆4], Cheng-lin Liu[†4] and Jean-Marc Ogier[⋆4]

⋆: L3i Laboratory, University of La Rochelle, France

†: National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences, China

•: Computer Vision Center, Universitat Autònoma de Barcelona, Spain

◇: Samsung R&D Institute of China, Beijing

⊙: Digital Media & Communications R&D Center Samsung Electronics, Seoul, Korea

‡: CVPR unit, Indian Statistical Institute, India

Corresponding author: nibal.nayef@univ-lr.fr

[1,2,3,4]: authors contributed equally as first, second, third or fourth author

***Abstract*—Text detection and recognition in a natural environment are key components of many applications, ranging from business card digitization to shop indexation in a street. This competition aims at assessing the ability of state-of-the-art methods to detect Multi-Lingual Text (MLT) in scene images, such as in contents gathered from the Internet media and in modern cities where multiple cultures live and communicate together. This competition is an extension of the Robust Reading Competition (RRC) which has been held since 2003 both in ICDAR and in an online context. The proposed competition is presented as a new challenge of the RRC. The dataset built for this challenge largely extends the previous RRC editions in many aspects: the multi-lingual text, the size of the dataset, the multi-oriented text, the wide variety of scenes. The dataset is comprised of 18,000 images which contain text belonging to 9 languages. The challenge is comprised of three tasks related to text detection and script classification. We have received a total of 16 participations from the research and industrial communities. This paper presents the dataset, the tasks and the findings of this RRC-MLT challenge.**

***Keywords*—Scene Text Detection, Multi-lingual Text, Script Identification**

I. Introduction and Related Work

In this RRC-MLT challenge, we try to answer the question whether text detection methods (whether deep learning-based or otherwise) could handle different scripts without fundamental changes in the used algorithms and techniques, or do we really need script-specific methods? The ultimate goal of robust reading is be able to read the text which appears in any captured image despite image source (type), image quality, text script or any other difficulties. Many research works have been devoted to solve this problem. The previous editions of RRC competitions [1], [2] and other works [3], [4], [5], [6], [7], have provided useful datasets to help researchers tackle each of those problems in order to robustly read text in natural scene images. In this challenge, we extend state-of-the-art work further by tackling the problem of multi-lingual text detection and script identification. In other words, participating methods should be script-robust text detection methods.

Despite the available datasets related to scene text detection or to script identification [2], [3], [4], [5], [6], [7], our dataset offers interesting novel aspects. The dataset is composed of widely variable scene images which contain text of one or more of 9 languages representing 6 different scripts. Our dataset contains many more images than related datasets (18,000 images). The number of images per script is equal. This makes it a useful benchmark for the task of multi-lingual scene text detection. The dataset along with its ground truth contains all necessary information to prepare for text recognition systems as well. The considered languages are the following: Arabic, Bangla, Chinese, English, French, German, Italian, Japanese and Korean.

Such dataset is the natural extension of the RRC series, with more scripts and more images while only considering intentional (focused) text. It addresses the needs of the research and industrial communities for improved and robust scene text detection and script classification. All the details about the RRC-MLT challenge and its dataset are available on the RRC competition website: http://rrc.cvc.uab.es/?ch=8.

The content of this paper is organized as follows. We first introduce the datasets used for the three tasks of the RRC-MLT competition (Sec. II), then, for each task, we describe its goal, its evaluation protocol, list the participant methods and discuss the results obtained by participants (Sec. III, IV, V). We conclude the paper and discuss future work in Sec. VI.

II. The "RRC-MLT" Dataset and Challenge Organization

A. The MLT Dataset

1) Type/source of images: The dataset is comprised of natural scene images with embedded text, such as street signs, street advertisement boards, shops names, passing vehicles and users photos in microblogs. The images have been mostly

captured by different users, using different mobile phone cameras. Some images have been carefully collected by Internet search so that they are similar to the captured ones and freely available. This kind of images represents one of the mostly encountered image types on the Internet which are the images with embedded text in social media. A large percentage of the images contain more than one language and/or script, such as in the MSRA-TD500[1] and KAIST[2] datasets. The key common aspect in these datasets is the wildness of text.

2) Homogeneity of the Dataset: We have imposed conditions on the collection of our dataset related to the type, contents and capture conditions of images. This is to ensure – to some extent – the homogeneity of the collected images of different scripts, since the images have been collected in different countries by different people. This makes our dataset a meaningful benchmark for judging the ability of algorithms to distinguish different scripts based on the characteristics of the scripts rather than the characteristics and patterns of the images containing the scripts.

3) Number of Images, Languages and Scripts: The dataset is comprised of 18,000 images containing text of 9 different languages. We have collected 2,000 images per language, however, an image could contain text of more than one language. This means, each language is represented in at least 2,000 images. The nine languages are: Arabic, Bangla, Chinese, English, French, German, Italian, Japanese and Korean. Those languages belong to one of the following six scripts: Arabic, Bangla, Chinese, Japanese, Korean and Latin. We have also added a script class called "Symbols" for text characters such as + / > :) ′ . " −. Such symbols are common in the different languages. Another additional script is called "Mixed" used for text of two or more scripts in one word (without spaces). The images will be divided as follows: 50% for training (a total of 9,000 images, 1,000 per language), and 50% for testing.

B. MLT Challenge Organization

This challenge is comprised of three tasks related to text detection and script classification (see Sections III, IV and V). We have used the web portal of the RRC platform[3] [2] for interacting with participants regarding the challenge schedule, downloads, online submissions, etc. The challenge for both training and test periods ran between 1st April to 1st July, where the test period started on 1st June (lasted one month). The participants have been asked to submit the results of their methods online according to specific formats.

Overall, we had 16 different participations distributed as follows: 6 in Task-1, 8 in Task-2 and 2 participants in Task-3. Some participants submitted results for more than one task using different methods. Note also that some of the participants have submitted multiple ***similar*** methods for the ***same*** task. In the cases where the submitted methods are not demonstrably different (and their descriptions clearly indicate this), the participants have been asked to choose one method – without knowing the results – to be published in this report, while the rest of their methods will be shown online on the RRC website. The participants and the descriptions of their methods are listed in *no particular order* within the sections that refer to the tasks in which they participated.

[1]http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)

[2]http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database

[3]http://rrc.cvc.uab.es/

III. Task-1: Multi-Lingual Text Detection

A. Task-1 Description

This task consists of detecting multi-lingual text at word level in a natural scene with focused text. A text word is defined as a consecutive set of characters without spaces, i.e. words are separated by spaces except in Chinese and Japanese where the text is labeled at line level.

In order to prepare their methods, participants were provided with a training set of 9,000 images along with their associated Ground Truth (GT). Each image has a corresponding GT file that contains a list of bounding boxes coordinates for each text word in the image and the transcription (though not used) of that word. Bounding boxes are represented by four corner points, allowing for any quadrilateral shape to be represented. In the ground truth, the text regions which were not readable by the annotators – due to low resolution or other distortions – are marked as "don't care" and ignored in the evaluation process.

The test set of this task consists in 9,000 images of full scene images. For each image in the test set, participants were expected to produce a list comprised of a four-corner bounding box for each word detected in the image.

B. Evaluation Protocol for Task-1

The f-measure (Hmean) is used as the metric for ranking the participants methods. The standard f-measure is based on both the recall and precision of the detected word bounding boxes as compared to the ground truth. A detection is considered as correct (true positive) if the detected bounding box has more than 50% overlap (intersection over union) with the GT box. The details of the evaluation are described in the following.

The f-measure is computed image by image and then aggregated. At the image level, the evaluation procedure works as follows: let $D = \{d_1, d_2, \ldots, d_k, \ldots, d_l\}$ be the set of bounding boxes of the "don't care" regions, $G = \{g_1, g_2, \ldots, g_i, \ldots, g_m\}$ be the set of bounding boxes in the ground truth, and $T = \{t_1, t_2, \ldots, t_j, \ldots, t_n\}$ be the set of bounding boxes in the results under evaluation.

First, the result bounding boxes from T are matched against the "don't care" regions set D to eliminate noise. Each quadrilateral t_j is compared against each quadrilateral d_k and t_j is discarded if the following condition is true:

$$area(d_k) = 0 \vee \frac{area(d_k) \cap area(t_k)}{area(d_k)} > 0.5 \qquad (1)$$

Such approach leads to some minor issues with ground truth regions overlapping with "don't care" regions. This was not anticipated in the previous version of the RRC evaluation platform. However, only few cases in the dataset were observed, and there was no impact on the global evaluation

of the methods. While this highlights possible improvement of the evaluation method in the future, the original evaluation method was used in its current state.

Once the set of the detected bounding boxes T is filtered, the resulting filtered set $T' = \{t'_1, t'_2, \ldots, t'_j, \ldots, t'_{n'}\}$ is matched against the set of ground truth quadrilaterals G. A positive match is counted each time a couple of elements (g_i, t'_j) verifies the following condition:

$$\frac{area(g_i) \cap area(t'_j)}{area(g_i) \cup area(t'_j)} > 0.5 \tag{2}$$

with $g_i \in G$ and $t'_j \in T'$. An extra test ensures that each element g_i and each element t'_j can only be matched once.

Given the set of positive (relevant) matches M, the set of expected words G and the set of filtered results T', we can compute the precision, recall and f-measure (harmonic mean of the precision and the recall) as follows:

$$\begin{aligned} precision &= \frac{|M|}{|T'|} \\ recall &= \frac{|M|}{|G|} \\ fmeasure &= \frac{2 \cdot precision \cdot recall}{precision + recall} \end{aligned} \tag{3}$$

When $T' = \emptyset$, then the precision is set to 0, as there always is at least one expected result in each image ($G \neq \emptyset$). Finally, the scores of all the images of the test set are averaged to produce the global score of a given method.

C. Participant Methods for Task-1

1) Team SCUT–DLVC Lab:
Authors: Yuliang Liu, Lianwen Jin, Sheng Zhang, Canjie Luo, Zhaohai Li, Lele Xie, Zenghui Sun
Affiliation: South China University of Technology
Method name: SCUT_DLVClab1
Method description: Two deep learning-based stages have been used, the first is to roughly detect text, and the second is to finely adjust the bounding box for accurate detection. A novel network is designed for multi-scale learning to generate tight quadrilateral proposals for the scene text. This is followed by a post-processing pipeline to improve the precision.

2) CAS & Fudan University:
Authors: Jianqi Ma, Weiyuan Shao, Yingbin Zheng, Hong Wang, Li Wang, Hao Ye, Xiangyang Xue
Affiliation: Shanghai Advanced Research Institute, CAS & Fudan University
Method name: SARI_FDU_RRPN_v1
Method description: A Rotation Region Proposal Network (RRPN) [8] is designed to generate inclined proposals with text orientation angle information. The angle information is used for bounding box regression to detect accurately oriented text proposals. The rotated region-of-interest pooling layer projects the proposals to a feature map for a text region classifier.

3) Sensetime OCR - A Deformable-FCN based network:
Authors: Xuebo Liu, Ding Liang, Dagui Chen, Minghao Guo, Junjie Yan
Affiliation: Sensetime Company - China
Method name: A Deformable-FCN based network
Method description: An FCN-based method is used where for every pixel in a image, the method predicts whether it's text, and also the text's location if it's in a text box. Online hard example mining and deformable convolution are used to improve the performance.

4) TH-DL:
Authors: Donglai Xiang, Jiaming Guo, Liangrui Peng, Changsong Liu
Affiliation: Tsinghua University, Beijing, China
Method name: TH-DL
Method description: A CNN with a multi-level feature pyramid is used. It consists of a modified FCN with residual connection as a proposal generator and a Fast R-CNN detector with Rotation RoI pooling for multi-oriented text detection. Firstly, an image is input into the FCN with residual connection which predicts a salient map that contains the probability of every pixel belonging to a text region. Then, the map is binarized at multiple thresholds, and connected components (CCs) are extracted. The CCs that break into multiple parts at a higher threshold are selected and their bounding boxes represent region proposals. Next, the features of the region proposals after Rotation RoI pooling are input into the Fast R-CNN network that filters non-text regions and regresses the bounding boxes to more accurate positions. Finally, non-maximum suppression is performed to obtain the text detection results.

5) Linkage-ER-Flow2017:
Authors: CVMT_text group
Affiliation: Computer and Control Engineering, University of Chinese Academy of Sciences
Method name: Linkage-ER-Flow2017
Method description: In this method, firstly, Extremal Regions (ERs) are robustly extracted from each channel of a given image. Secondly, the linkage between the components is calculated and a pruning method of repeated Connected Components (CCs) is used to get a set of non-overlapping CCs. Minimum cost maximum flow is used to extract candidate lines of two directions, and a merging strategy is used to combine results of all directions and channels. Non-text lines with a small average CNN score are eliminated.

6) Alibaba group:
Authors: Enhua Cao, Dong Yi, Fan Zhang and Rufeng Chu
Affiliation: Alibaba group
Method name: IDST_CV
Method description: A hybrid strategy to detect multi-lingual text is used. The method includes bottom-up and top-down steps. Firstly, a CNN is used to detect local text regions and predict their relationships. Based on the outputs of the CNN, the local regions are robustly grouped into text lines. Finally, a language-sensitive CNN+RNN network is trained to learn how to divide a text line into words.

D. Results of Task-1

We report here the results obtained by the participants for this first task. The ranking of the participants according to f-measure is summarized in Table I.

The results of the top three methods are close. All the methods except the second method have been tuned to get a better precision than recall. The results show that the task is challenging to participants compared to similar tasks of the previous RRC editions. This is due not only to the multi-script

TABLE I. RESULTS OF THE RRC-MLT CHALLENGE FOR TASK-1: MULTI-LINGUAL TEXT DETECTION

Ranking	Method	F-Measure	Recall	Precision
1	SCUT_DLVClab	64.96%	54.54%	80.28%
2	Sensetime OCR	62.56%	69.43%	56.93%
3	SARI_FDU_RRPN_v1	62.37%	55.50%	71.17%
4	TH-DL	45.97%	34.78%	67.75%
5	linkage-ER-Flow	32.49%	25.59%	44.48%
6	IDST_CV	28.63%	26.02%	31.81%
Baseline	NLPR-PAL [9]	66.01%	57.94%	76.69%

text, but also to the large dataset size and the huge variety in the dataset images in terms of both content and resolution. The high-performing methods have used recent techniques in deep learning such as region proposals, multi-stage detection networks, bounding box regression and deformable convolution. The baseline method at the bottom of Table I has been submitted by the organizers as a baseline after the end of the competition, and hence, we do not rank it.

IV. TASK-2: CROPPED WORD SCRIPT IDENTIFICATION

A. Task Description

The text in our dataset images appears in 9 different languages, some of them share the same script. Additionally, punctuation and some math symbols sometimes appear as separate words, those words are annotated as a special script class called "Symbols". Hence, we have a total of 7 different scripts. We have excluded the words that have "Mixed" script for this task due to the very small number of samples. We have also excluded all the "don't care" words whether they have a recognizable script or not.

The training and test sets of this task consist of isolated word images that have been extracted from the full images of Task-1. Each word image is associated with a script class in the ground truth. In total, there are 84,868 training images and 97,619 testing images.

For this task, we provide all the words in our dataset in separate image files, along with the corresponding ground truth script *ID* and the transcription of each word image. The transcription is not used in this task. For each text block, the axis-oriented bounding box that tightly contains the text block is provided.

As the task for participants, we provide the cropped images of text words and we ask for the script *ID* of each image. A single script name (*ID*) per image is requested. The valid scripts for this task are: "Arabic", "Bangla", "Chinese", "Japanese", "Korean", "Latin" and "Symbols".

B. Evaluation Protocol for Task-2

The evaluation of results against the ground truth is computed in the following way: participants provide a script *ID* for each word image, and if the result is correct, then the count of correct results is incremented. The final evaluation for a given method is the accuracy of such prediction. This can be summarized with the simple definition that follows: let $G = \{g_1, g_2, \ldots, g_i, \ldots, g_m\}$ be the set of correct script classes in the ground truth, and $T = \{t_1, t_2, \ldots, t_i, \ldots, t_m\}$ be the set of script classes returned by a given method, where g_i and t_i refer to the same original image. Then, the performance of a given method is expressed by:

$$accuracy = \frac{1}{m} \sum_{i=1,\ldots,m} \begin{cases} 1 & \text{if } g_i = t_i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

C. Participant Methods for Task-2

1) Team SCUT–DLVC Lab:
Authors: Canjie Luo, Zhaohai Li, Lianwen Jin, Zenghui Sun, Yuliang Liu, Qinghe Zeng
Affiliation: South China University of Technology
Method name: SCUTDLVClab
Method description: A CNN-based classification method is used. During the training phase, sub-group-of-images are randomly cropped. In the test phase, a novel sliding window method is applied on the entire image, which can be regarded as convolutional stride (replaced full connection layer by convolution layer). The category with the top confidence is chosen as the final result. A image-size normalization method is also used for further improving the results.

2) CNN-based method:
Authors: Yash Patel[1,2], Michal Buta[1], Luk Neumann[1], Jiri Matas[1]
Affiliations: [1] Centre for Machine Perception, Department of Cybernetics, Czech Technical University, Prague, Czech Republic, [2] CVIT, KCIS, IIIT Hyderabad, India.
Method name: CNN-based method
Method description: A CNN-based approach is used for script-identification in cropped word images. The convolutional layers from VGG-16 architecture are used along with a Global-Average-Pooling and two fully connected layers. To preserve the aspect ratio of input images in both training and testing, the images are resized into fixed-height (64) and variable-width tensors. For training, the convolutional layers are initialized with ImageNet weights. The categorical-cross-entropy loss is utilized, and all the layers (both convolutional and fully connected) are updated during back-propagation.

3) TNet:
Authors: FlyText team
Affiliation: CVMT lab, University of Chinese Academy of Sciences
Method name: TNet
Method description: A deep network is used for classification. Firstly, a joint strategy is used to enhance the features of the dataset, and then the deep network is used for training. Then, the majority vote is used to determine the script class.

4) BLCT:
Authors: Jan Zdenek, Hideki Nakayama
Affiliation: The University of Tokyo, Graduate School of Information Science and Technology
Method name: BLCT
Method description: A CNN is combined with the bag-of-visual-words approach. A patch-based approach is adopted to solve the issue of variable sizes and aspect ratios of the input images. Individual local patches extracted from training image data are used to train the CNN with 6 convolutional layers. Feature vectors of all patches from each training image are fed to the trained CNN and the output is extracted from the penultimate layer of the network. Random combinations

of feature vectors are created to form local convolutional triplets and the 3 vectors in each triplet are added. The local convolutional triplets are used to create a bag-of-visual-words vocabulary with the size of 1024 codewords. Each image is then represented as a vector of codewords which are then aggregated into histograms of occurrences. The histograms are used for global representation of each image. An MLP with two hidden layers and a "Dropout" after each layer is used for the final classification.

5) TH-DL:
Authors: Jiaming Guo, Guangxiang Bin, Liangrui Peng
Affiliation: Tsinghua University, Beijing, China
Method name: TH-DL
Method description: A deep CNN similar to GoogLeNet is used with a smaller number of layers of inception structures for computation efficiency. For image pre-processing, the shorter edge is resized to 224 while preserving the aspect ratio of the original image. Average pooling is used to transform the spatial dimension of the feature map into a fixed size before the final fully connected layer. In the training process, the batch size for each iteration is set to 1, the mean of gradients for a preset size of iterations (e.g. 32) are calculated and used to update the network weights.

6) TH-CNN:
Authors: Yejun Tang, Haoyu Qin, Liangrui Peng
Affiliation: Department of Electronic Engineering, Tsinghua University, Beijing, China
Method name: TH-CNN
Method description: A simplified GoogLeNet is used (Caffe implementation). The network is trained by using augmented samples. The original samples in the training set are rotated, blurred, mirrored and inverted. The numbers of training samples of different scripts are balanced. The input images are resized into 256x256 pixels and cropped into 227x227 pixels.

7) Synthetic-ECN:
Authors: Yizhi Wang
Affiliation: Peking University
Method name: A Synthetic-ecn method
Method description: The ECN-based model [10] is used. Firstly, large numbers of text images are generated in different scripts to ensure a balanced distribution in training data. Then the ECN-based model [10] is used to train a classification model.

8) An approach towards Word-Level Multi-Script Ident...:
Authors: Arindam Das, Saikat Roy
Affiliation: Imaging Tech Lab, HCL Technologies, Chennai, India.
Method name: An approach towards Word-Level Multi-Script Identification using Deep Transfer Features and SVM
Method description: A pre-trained model of VGG16 is used where weights are adapted the problem of script identification. Each labeled image is initially resized to 224x224 and passed through this deep CNN as a 3D matrix to extract features. The images in each set are first normalized based on mean and standard deviation of the training set. The CNN was not trained further, but the features (4096 sized vectors) are extracted from the last fully connected layer through forward propagation (for each dataset). An SVM with RBF Kernel is used as classifier and trained on the training set. An accuracy of 85.03% was achieved on the validation set, the same hyper-parameters are used to predict the scripts in the test set.

TABLE II. RESULTS OF THE RRC-MLT CHALLENGE FOR TASK-2: CROPPED WORD SCRIPT IDENTIFICATION

Ranking	Method	Accuracy
1	CNN based method	88.09%
2	SCUT-DLVClab	87.69%
3	BLCT	86.34%
4	TH-DL	80.72%
5	synthetic-ecn method	79.20%
6	An approach towards Word-Level Multi-Script Ident...	74.81%
7	TNet	48.33%
8	TH-CNN	43.22%

D. Results of Task-2

We report here the results obtained by the participants for this second task. The ranking of the participants – according to script classification accuracy – is summarized in Table II.

Most of the methods – including the top three – are based on a CNN approach. Based on the confusion matrices obtained for each method for the 7 scripts, we note that the top sources of errors were: (1) confusing Chinese and Japanese scripts and (2) classifying Japanese and Korean scripts as Latin (much more than the opposite). The first type of error is due to the similarity between one type of Japanese scripts (Kanji) and the Chinese script. One of the reasons for the second error type is the appearance of numbers (Latin script) within words of Japanese, Korean and Chinese scripts.

V. TASK-3: JOINT TEXT DETECTION AND SCRIPT IDENTIFICATION

A. Task Description

This task combines all the preparation steps needed for multi-lingual text recognition. The input of this task is a full scene image, and the output is a list of the bounding boxes of all the words in the image and the associated script *ID* for each word.

Similar to Task-1, the training and test sets are comprised each of 9,000 images which are the same images described in Task-1. The ground truth of the training set given to participants contains the coordinates of the bounding boxes of all the words inside an image (including "don't care" words), the transcription and the script *ID* for each text box.

For the participants, the first part of the required output is also similar to Task-1, namely the list of the detected bounding boxes. A second output part is required for this task, where for each detected bounding box (word) in the list, the script class has to be identified.

B. Evaluation Protocol for Task-3

The evaluation of this task is a cascade of correct localization (detection) of a text box ***and*** correct script classification. In practice, it only requires to inject the control of the correct identification of the script for a given text region into Eq. 2:

$$\frac{area(g_i) \cap area(t'_j)}{area(g_i) \cup area(t'_j)} > 0.5 \wedge scriptid(t'_j) = scriptid(g_i) \quad (5)$$

TABLE III. RESULTS OF THE RRC-MLT CHALLENGE FOR TASK-3: JOINT TEXT DETECTION AND SCRIPT IDENTIFICATION

Ranking	Method	F-Measure	Recall	Precision
1	SCUT-DLVClab2	58.08%	48.77%	71.78%
2	TH-DL	39.37%	29.65%	58.58%

C. Participant Methods for Task-3

1) TH-DL:
Authors: Donglai Xiang, Jiaming Guo, Guangxiang Bin, Liangrui Peng, Changsong Liu
Affiliation: Tsinghua University, Beijing, China
Method name: TH-DL
Method description: The method is an integration of the methods of Tasks 1 and 2 (see their descriptions above under the same method name). All the methods are implemented by PyTorch.

2) SCUT-DLVClab:
Authors: Yuliang Liu, Canjie Luo, Lianwen Jin, Sheng Zhang, Zhaohai Li, Lele Xie, Zenghui Sun
Affiliation: South China University of Technology
Method name: SCUT-DLVClab
Method description: Two models have been trained separately: one model for text detection and another for classifying scripts. The two models are jointed to output the final results. After generating the detection results, a classification model with 8 classes (including background) is used to discard the detected boxes classified as background with very high confidence. Then a 7-class model was utilized to yield the final results. Since Chinese and Japanese are found to be easily confused by the model, and since only few images simultaneously contain both Chinese and Japanese scripts, a statistical average method is used to modify the Chinese and Japanese classification results.

D. Results of Task-3

As this task requires a method that combines both detection and script classification, we had only two participant methods. We report here the results obtained by the participants for this third task. The ranking of the participants is summarized in Table III.

The results of the first method – the winning method – are comparable to the results achieved in Task-1 by the same participant team. The lower f-measure value is due to the errors in script classification. We note also that the methods have been tuned to achieve higher precision than recall. Overall, based on the results of both Task-1 and Task-3, we conclude that the dataset is very challenging, but not only due to the multi-script text. Other main sources of difficulty are the high variety in image content and the great difference in the resolutions of the images. Despite the large number of images in the dataset, there could still be a need to have more training samples that correspond in content and resolution to test images.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

This report summarizes the organization and the findings of the ICDAR2017 MLT challenge of the RRC competition. There has been a total of 16 participations in the 3 proposed tasks. This shows an interest of the community in the problem of multi-lingual scene text detection.

Our work extends the previous RRC editions in different aspects: the size of the dataset, the multi-lingual text, multi-oriented text, the wide variety of scenes in terms of content and image resolution. All the details about the RRC-MLT challenge and its dataset are available on the RRC competition website: http://rrc.cvc.uab.es/?ch=8.

Future versions of this challenge could tackle problems such as very large-scale multi-lingual scene text detection, more languages (of similar and also of different scripts), unfocused scene text and video text. This work provides the base on which such extensions could be built.

ACKNOWLEDGMENTS

This work is partially funded by Agence Nationale de la Recherche (ANR) in France and National Natural Science Foundation of China (NSFC 61411136002) in China under the AUDINM project.

REFERENCES

[1] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *The 12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 1484–1493.

[2] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading," in *The 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1156–1160.

[3] B. Shi, X. Bai, and C. Yao, "Script identification in the wild via discriminative convolutional neural network," *Pattern Recognition*, vol. 52, pp. 448–458, 2016.

[4] N. Sharma, R. Mandal, R. Sharma, U. Pal, and M. Blumenstein, "ICDAR2015 competition on video script identification (cvsi 2015)," in *The 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1196–1200.

[5] A. K. Singh, A. Mishra, P. Dabral, and C. V. Jawahar, "A simple and effective solution for script identification in the wild," in *Workshop on Document Analysis Systems (DAS)*, 2016, pp. 428–433.

[6] L. G. i Bigorda and D. Karatzas, "A fine-grained approach to scene text script identification," in *Workshop on Document Analysis Systems (DAS)*, 2016, pp. 192–197.

[7] D. Kumar, M. N. A. Prasad, and A. G. Ramakrishnan, "Multi-script robust reading competition in ICDAR 2013," in *The 4th International Workshop on Multilingual OCR*, ser. MOCR '13, 2013, pp. 14:1–14:5.

[8] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *CoRR*, vol. abs/1703.01086, 2017.

[9] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," *arXiv preprint arXiv:1703.08289*, 2017.

[10] L. G. i Bigorda, A. Nicolaou, and D. Karatzas, "Improving patch-based scene text script identification with ensembles of conjoined networks," *Pattern Recognition*, vol. 67, pp. 85–96, 2017.