ICDAR 2015 Competition on Robust Reading

Dimosthenis Karatzas[1], Lluis Gomez-Bigorda[1], Anguelos Nicolaou[1], Suman Ghosh[1], Andrew Bagdanov[1],
Masakazu Iwamura[2], Jiri Matas[3], Lukas Neumann[3], Vijay Ramaseshan Chandrasekhar[4],
Shijian Lu[4], Faisal Shafait[5], Seiichi Uchida[6], Ernest Valveny[1]
[1]Computer Vision Centre, Universitat Autònoma de Barcelona; {*dimos, lgomez, anguelos, sghosh, bagdanov, ernest*}*@cvc.uab.es*
[2]Osaka Prefecture University, Japan; *masa@cs.osakafu-u.ac.jp*
[3]Czech Technical University in Prague, Czech Republic; {*matas, neumalu1*}*@cmp.felk.cvut.cz*
[4]Institute for Infocomm Research, Singapore; {*vijay, slu*}*@i2r.a-star.edu.sg*
[5]National University of Science and Technology (NUST), Pakistan; *faisal.shafait@seecs.nust.edu.pk*
[6]Kyushu University, Japan; *uchida@ait.kyushu-u.ac.jp*

***Abstract*—Results of the ICDAR 2015 Robust Reading Competition are presented. A new Challenge 4 on Incidental Scene Text has been added to the Challenges on Born-Digital Images, Focused Scene Images and Video Text. Challenge 4 is run on a newly acquired dataset of 1,670 images evaluating Text Localisation, Word Recognition and End-to-End pipelines. In addition, the dataset for Challenge 3 on Video Text has been substantially updated with more video sequences and more accurate ground truth data. Finally, tasks assessing End-to-End system performance have been introduced to all Challenges. The competition took place in the first quarter of 2015, and received a total of 44 submissions. Only the tasks newly introduced in 2015 are reported on. The datasets, the ground truth specification and the evaluation protocols are presented together with the results and a brief summary of the participating methods.**

I. Introduction

Robust Reading refers to the automatic interpretation of written communication in unconstrained settingssuch as born-digital and real scene images and videos. The Robust Reading Competitions series addresses the need to quantify and track progress in this domain. The competition dates back to 2003[1] [2] [3], and was substantially revised in 2011 and 2013 [4] [5] [6], creating a comprehensive reference framework for robust reading pipelines evaluation [7]. The competition was open in a continuous mode between editions, allowing the submission and evaluation of results at any time. This has led to the acceptance of the Robust Reading Competition framework by researchers worldwide as the de-facto standard for evaluation, and has promoted good practice in the field. Over the past 1.5 year, the Web portal of the competition has been visited more than 140,000 times while about 2,000 results submissions from more than 750 registered users have been received and processed.

The 2015 edition of competition brings major changes. First, a new challenge on Incidental Scene Text (Challenge 4) is introduced, based on a new dataset of 1,670 images (17,548 annotated regions) acquired using the Google Glass. *Incidental Scene Text* refers to text that appears in the scene without the user having taken any prior action to cause its appearance in the field of view, or improve its positioning or quality in the frame. While focused scene text (Challenge 2) is the expected input for applications such as translation on demand, incidental scene text covers another wide range of applications linked to wearable cameras or massive urban captures where the capture is difficult or undesirable to control.

TABLE I. Evolution of the Robust Reading Competition.

Tasks	Challenge 1: Born-Digital	Challenge 2: Focused Scene Text	Challenge 3: Text in Videos	Challenge 4: Incidental Scene Text
1. Localization	2011 / 2013	2011 / 2013	2013 / 2015 (Table VI)	2015 (Table III)
2. Segmentation	2011 / 2013	2013		
3. Recognition	2011 / 2013	2011 / 2013		2015 (Table IV)
4. End-to-End	2015 (Table VIII)	2015 (Table IX)	2015 (Table VII)	2015 (Table V)

Second, tasks assessing "End-to-End" system performace have been introduced in all competition Challenges. The objective is to simultaneously localise and recognise of all words in the image or video sequence, modelling complete systems for text understanding.

Finally, the datasets for Challenge 3 on Video Text have been substantially updated, bringing the number of sequences up to 49, comprising a total of 27,824 frames (184,687 annotated regions).

In addition, the 2015 edition offers improved and intuitive performance evaluation protocols, available through the Web portal[1] that allows the continuous submission of new methods, on-line performance evaluation and enhanced visualisation.

II. Competition Organisation

An overview of the evolution of the Robust Reading Competition is given in Table (I). The competition is organised around four Challenges, each based on a series of specific tasks. The highlight of the 2015 competition, and coverage of this report, is on the newly introduced tasks, highlighted in Table I. Older tasks are covered in previous reports [6][5][4] while up to date information is available on the competition Web site.

The Competition run between January and April 2015, in open mode, meaning that the results were provided by authors themselves and the data were public. The authors were allowed to make multiple submissions to the same task as long as their submissions reflected sufficiently different pipelines. Submissions based on the same pipeline (typically reflecting different parameter configurations), were filtered considering only the latest submission as a valid competition entry . In total, 44 submissions (out of 76) were accepted as valid, after filtering multiple submissions based on the same pipeline. The

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

978-1-4799-1805-8/15/$31.00 ©2015 IEEE

TABLE II. RRC 2015 PARTICIPANTS

"Method", Authors, *Affiliation*	Challenge Task						
	1.4	2.4	3.1	3.4	4.1	4.3	4.4
"DSM" S. Kim, *Qualcomm*						●	
"AJOU" H. I. Koo, and Y. G. Kim, *Ajou University* [8, 9]			●		●		
"Beam Search CUNI" J. Libovický, and P. Pecina, *Charles University in Prague*		●					●
"Beam Search CUNI +S" –//–		●					●
"Deep2Text-I" X. C. Yin, C. Yang, J. B. Hou, W. Y. Pei, X. Yin, and K. Huang, *University of Science and Technology Beijing and Xi'an Jiaotong-Liverpool University* [10, 11, 12, 13]	●	●	●	●			
"Deep2Text-II" –//–	●	●					
"Deep2Text-MO" X. C. Yin, W. Y. Pei, C. Yang, Y. Zheng, Q. Gao, G. Ji, and X. Yin, *University of Science and Technology Beijing*					●		●
"HUST MCLAB" B. Shi, C. Zhang, C. Yao, X. Bai, and Z. Zhang, *Huazhong University of Science and Technology*					●		
"MAPS" D. Kumar and A. G. Ramakrishnan, *Dayananda Sagar Institutions and Indian Institute of Science* [14]						●	
"NESP" –//– [15]						●	
"MSER MRF" X. Liu, *NanJing University*		●					
"NJU Text" F. Su, H. Xu, and T. Lu, *Nanjing University*	●	●			●		●
"PAL" Y. C. Wu, K. Chen, X. He, Z. Chen, F. Yin, and C. L. Liu, *CASIA NLPR* [16, 17]	●						
"RTST Lucas-Kanade-2" Y. Zhou, and H. Lai, *NJUCS Nanjing University* [18]			●				
"Stradvision-1" H. Cho, M. Sung, and B. Jun, *Stradvision*	●	●	●	●	●		●
"Stradvision-2" –//–	●				●		●
"TextCatcher-1" J. Fabrizio, M. Robert-Seidowsky, *LRDE*	●						
"TextCatcher-2" J. Fabrizio, M. Robert-Seidowsky, E. Carlinet, T. Geraud, *LRDE*	●				●		
"USTB-TexVideo" X. Yin, S. Tian, Z. Y. Zuo, W. Y. Pei, and C. Yang, *University of Science and Technology Beijing* [10, 11, 12]			●	●			
"USTB TexVideo-II-1" –//–			●	●			
"USTB texVideo-II-2" –//–			●	●			
"VGGMaxBBNet" A. Gupta, M. Jaderberg, A. Zisserman, *Visual Geometry Group, University of Oxford* [12]		●					
"CNN MSER" W. He, *CASIA*					●		

list of submitted methods is summarised in table II. Due to space limitations full descriptions of all participating methods can be found on the competition Web[2].

The presentation of this report is structured according to the novelties of the 2015 competition. The new Challenge 4 is presented in section III. Challenge 3, which has been substantially updated is covered in section IV. The End-to-End tasks introduced in Challenges 1 and 2 are covered in section V. Overall conclusions are presented in Section VI.

III. CHALLENGE 4: INCIDENTAL SCENE TEXT

Challenge 4 focuses on real scene images. Unlike Challenge 2 which is based on well-captured images focusing on the text content, Challenge 4 addresses *incidental* scene text. Incidental scene text refers to text that appears in the scene without the user having taken any prior action to cause its appearance in the field of view, or improve its positioning / quality in the frame.

The dataset of Challenge 4 was collected over a period of a few months in Singapore.The focus of the current edition of the competition is on Latin-scripted text. The dataset also contains text in a number of Orient scripts, currently treated as *do not care* regions (see below).The ICDAR 2015 Incidental Scene Text dataset comprises 1,670 images and 17,548 annotated regions, making it one of the largest, public domain, fully ground truthed datasets available. 1,500 of the images have been made publicly available, split between a training set of 1,000 images and a test set of 500. The remaining 170 images comprise a sequestered, private set.

The dataset has been annotated through a collective international effort, involving 6 institutions worldwide. For this purpose we used the Web framework [3] developed for the Robust Reading Competition by the Computer Vision Centre [7].

The ground truth for Challenge 4 comprises word-level bounding boxes, along with their Unicode transcriptions. Word regions are defined by quadrilaterals, as opposed to axis-oriented rectangles. This is necessary in the case of incidental text, as perspective distortion can be significant. To ensure consistency in the definition of the word regions, a real time preview of a rectified view of the word region was provided, and ground-truthers were required to adjust the area so that the rectified word appears correct.

Word regions were classified as either *care* or *do not care*. *Do not care* words include text in non-Latin scripts, and text that the ground-truther deemed as non-readable. One- and two-character words are automatically marked as *do not care* regions. Performance evaluation is based only on the subset of *care* words, while the performance of a method over *do not care* words does not affect the results. During a second-pass verification, rectified word regions were presented in random order. This permits assessing word readability on its own, without being influenced by any textual or visual context.

In addition, a set of training and test vocabularies were provided. The use of controlled vocabularies defines some minimal common conditions for recognition that permit meaningful method comparisons. The vocabularies provided are:

- **Strongly Contextualised**: per-image vocabularies of 100 words comprising all words in the corresponding image as well as distractor words selected from the rest of the training/test set (see Wang et al. [19])
- **Weakly Contextualised**: a vocabulary of all words in the training/test set
- **Generic**: a generic vocabulary of about 90K words derived from the dataset[4] of Jaderberg et al. [12]

The vocabularies provided exclude words of one or two characters and do not contain alphanumeric structures such as prices, URLs, dates etc. If such structures were to be included in a vocabulary they should rather be defined as regular expressions and not explicitly. Nevertheless, such structures (instantiations of the corresponding regular expressions) are tagged in the images and a good recognition method is expected to recognise them.Words were stripped by any preceding or trailing characters other than the letters of common Latin scripts before they were added in the vocabulary. This includes punctuation marks, numerical and other symbols. See the Web of the competition for more details.

A. TASK 4.1: Text Localisation of Incidental Scene Text

The objective of this task is the correct localisation of all *care* words of the image. Performance evaluation is based on a single Intersection-over-Union criterion, with a threshold of 50%, in accordance to standard practice in object

[2] Available through http://rrc.cvc.uab.es

[3] The 2013 version is available from http://www.cvc.uab.es/apep/

[4] Available at: http://www.robots.ox.ac.uk/~vgg/data/text/

TABLE III. RANKING IN TASK 4.1 (INCIDENTAL TEXT LOCALISATION)

Method	Precision %	Recall %	F-Score %
Stradvision-2	**77.46**	36.74	**49.84**
Stradvision-1	53.39	46.27	49.57
NJU	70.44	36.25	47.87
AJOU[8]	47.26	**46.94**	47.1
HUST-MCLAB	44.0	37.79	40.66
Deep2Text-MO[10, 11]	49.59	32.11	38.98
CNN MSER	34.71	34.42	34.57
TextCatcher-2	24.91	34.81	29.04

TABLE IV. RANKING IN TASK 4.3 (INCIDENTAL TEXT RECOGNITION)

Method	Total Edit Distance	Correctly Recognised Words
MAPS[14]	**1128.0**	**32.93**
NESP[15]	1164.6	31.68
DSM	1178.8	25.85

recognition [20]. Using this framework, granularity differences between the ground truth and the detections are penalised. Any detections overlapping by more than 50% with *do not care* ground truth regions are filtered before evaluation takes place, while ground truth regions marked as *do not care* are not taken into account at the time of evaluation.

Seven methods were submitted to this task and their results are shown in Table III. Methods are ranked based on their *F-score*. All metrics are calculated cumulatively over the whole test set (all detections over all images pooled together).

In terms of Precision, almost all methods are below 50%, with the exception of "Stradvision-2" and "NJU", which yield precision values above 70%. A closer examination reveals that these methods made use of the vocabularies provided for Task 4 to filter localisation results. Although these dictionaries were not meant to be used for localisation, this was not explicitly forbidden. In terms of Recall, most methods perform below 40% with the exception of "AJOU" and "Stradvision-1", which are both based on variants of the MSER algorithm followed by different grouping approaches.

B. TASK 4.3: Word Recognition of Incidental Scene Text

This task aims to evaluate recognition performance over a set of pre-localised word regions. The dataset comprises axis-aligned cut-out regions of all *care* words in the corresponding subset, along with the quadrilateral coordinates defining the location of the word within the axis-aligned bounding box provided.

During test time the authors had access to all vocabularies provided while they were free to incorporate other vocabularies / text corpuses to enhance their language models. The evaluation protocol is based on a standard edit distance metric, with equal costs for additions, deletions and substitutions [6]. For each word we calculate the normalized edit distance to the length of the ground truth transcription. The comparison is case sensitive. Statistics on the percentage of correctly recognised words are also provided.

Three methods were submitted to this task, the results of which are shown in Table IV. The sum of normalised edit distances over all words of the test set was used to rank the methods. MAPS is the method that yields the smallest Total Edit Distance, although the performance of all three methods is very similar to allow any safe conclusions. On the other hand, it seems that the NESP and MAPS methods have a clear edge over the DSM method in terms of correctly recognised words. It transpires that NESP and MAPS make use of OmniPage OCR for recognition, instead of an in-house recogniser (DSM).

TABLE V. RANKING IN TASK 4.4 (INCIDENTAL TEXT END-TO-END)

Method	Strong			Weak			Generic		
	P(%)	R(%)	F	P(%)	R(%)	F	P(%)	R(%)	F
Stradvision-2	67.92	32.21	**43.7**	-	-	-	-	-	-
Baseline (TextSpotter)[21]	62.21	24.41	35.06	24.96	**16.56**	**19.91**	18.32	13.58	15.6
Stradvision-1	28.51	**39.77**	33.21	-	-	-	-	-	-
NJU	48.8	24.51	32.63	-	-	-	-	-	-
Beam Search CUNI	37.83	15.65	22.14	33.72	14.01	19.8	29.64	12.37	**17.46**
Deep2Text-MO[10, 11]	21.34	13.82	16.77	21.34	13.82	16.77	21.34	**13.82**	16.77
Baseline (OpenCv + Tessaract)[22]	40.9	8.33	13.84	32.48	7.37	12.01	19.3	5.06	8.01
Beam Search CUNI +S	**81.08**	7.22	13.26	**64.74**	5.92	10.85	**34.96**	3.8	6.86

C. TASK 4.4: End-to-End Systems for Incidental Scene Text

This task aims to assess End-to-End system performance. The evaluation strategy combines measuring localisation efficiency and recognition capacity over all *care* words.During testing, the authors could make use of the three types of vocabularies provided, defining three evaluation scenarios of increasing difficulty. Submitting results based on the strongly contextualised vocabulary was obligatory, while results based on the weakly contextualised and generic ones were optional.

Correct localisation was assessed in the same way as in Task 4.1 (see Section III-A). Subsequently, the recognition output for correctly localised words was compared to the ground truth transcription and a perfect match was sought. For this string comparison we do not take into account any punctuation marks at the beginning or the end of the word.

Two baselines based on public domain methods are given for this task. "Baseline (TextSpotter)" is an unconstrained real-time end-to-end text localization and recognition method [21]. The real-time performance is achieved by posing the character detection problem as an efficient sequential selection from the set of Extremal Regions (ERs). ERs are grouped into word regions which are recognized using an approximate nearest-neighbour classifier operating on a coarse Gaussian scale-space pyramid. A demo of the software is available online[5]. "Baseline (OpenCV + Tesseract)" makes use of the publicly available pipeline[6] proposed in [22]. Concretely, we use the OpenCV Class Specific Extremal Regions (CSER) and Exhaustive Search algorithms initially proposed by Neumann and Matas [18] along with the perceptual grouping approach of Gomez and Karatzas [23] for text localisation. Text recognition is performed using the open source Tesseract OCR engine[7].

Six methods were submitted to this task the results of which are shown in Table V. *F-score* was used for ranking. All metrics are calculated cumulatively over the whole test set. The "Stradvision-2" method yields the highest F-score using Strongly Contextualised dictionaries, without achieving top performance in neither Precision nor Recall. This method seems to be adding a post-processing step to "Stradvision-1", filtering words according to the provided vocabulary. "Stradvision-1" yields very low Precision, but the highest Recall values, and the post-processing step of "Stradvision-2" more than doubles Precision, while it moderately affects Recall. "Beam Search CUNI +S", yields the top score in Precision, but at the cost of the lowest Recall score. The method "Deep2Text-MO" does not seem to make use of the

[5] http://www.textspotter.org

[6] https://github.com/ComputerVisionCentre/RRC2015_Baseline_CV3Tess

[7] https://code.google.com/p/tesseract-ocr/

TABLE VI. RANKING IN TASK 3.1 (VIDEO TEXT LOCALISATION)

Method	MOTP	MOTA	ATA
Deep2Text-I[10, 11]	71.01	40.77	**45.18**
USTB-TexVideo[10, 11]	71.33	49.33	41.31
AJOU[8]	**73.25**	**53.45**	38.77
USTB-texVideo-II-2[10, 11]	72.47	50.38	35.71
Stradvision-1	70.82	47.58	32.12
USTB-TexVideo-II-1[10, 11]	69.51	19.69	30.15
RTST-LucasKanade-2	64.44	-20.28	0.34

TABLE VII. RANKING IN TASK 3.4 (VIDEO TEXT END-TO-END)

Method	MOTP	MOTA	ATA
Baseline (TextSpotter)[21]	**69.51**	**59.83**	**41.84**
Stradivision-1	69.21	56.54	28.53
USTB-TexVideo[10, 11]	65.08	45.82	19.85
Deep2Text-I[10, 11]	62.12	35.39	18.64
USTB-texVideo-II-2[10, 11]	63.48	50.52	17.8
USTB-TexVideo-II-1[10, 11]	60.46	21.16	13.79

provided vocabularies as its performance remains the same for all levels of vocabulary contextualisation.

IV. CHALLENGE 3: READING TEXT IN VIDEOS

Challenge 3 on Text in Videos evaluates the use of temporal information to improve text detection and recognition performance. The dataset has been substantially updated with new scene video sequences, resulting to a training set of 25 videos (13,450 frames) and a test set of 24 videos (14,374 frames). The ground truth quality has been improved at the frame and sequence level, while a new task assessing End-to-End performance has been introduced.

A. *TASK 3.1: Text Localisation in Video*

The task requires that words are both detected and tracked correctly over the video sequence. The evaluation is based on an adaptation of the CLEAR-MOT framework [24] for multiple object tracking. For each method we provide three different metrics: the Multiple Object Tracking Precision (MOTP), the Multiple Object Tracking Accuracy (MOTA), and the Average Tracking Accuracy (ATA). See the 2013 competition report [6] for details about these metrics.

Seven methods were submitted to this task, the results of which are shown in Table VI. The ranking metric is the ATA measure which summarizes the overall performance of methods over whole video sequences. All participating methods achieve an ATA measure under 50%, which indicates that text detection and tracking in this dataset is still very challenging for current state-of-the-art. The winning method for this task, "Deep2Text I (Video)", is an evolution of the winner in Task 2.1 of the last contest ("USTB_TexStar"). The "AJOU" method show the best numbers in MOTA and MOTP, indicating a better performance in terms of text detection, but the top performer, "Deep2Text I", has superior tracking performance, reflected by its a higher ATA by 6%.

B. *TASK 3.4: End-to-End Systems for Video Text*

This task requires that words that are correctly localised in every frame and correctly tracked over the video sequence are also correctly localised at the sequence level. The same dataset and ground truth as Task 3.1 is used. In addition to localisation and transcription ground truth, a series of vocabularies for the training and the test set are provided, similarly to Task 4.4 (see Section III-C).

TABLE VIII. RANKING IN TASK 1.4 (BORN DIGITAL END-TO-END)

Method	Strong			Weak			Generic		
	P(%)	R(%)	F	P(%)	R(%)	F	P(%)	R(%)	F
Stradvision-2	83.93	73.02	**78.1**	77.61	70.86	74.08	57.35	56.68	57.01
Deep2Text-II[10, 11]	80.97	**73.37**	76.98	80.97	**73.37**	**76.98**	80.97	**73.37**	**76.98**
Stradvision-1	**84.72**	70.17	76.76	78.9	67.87	72.97	58.2	54.31	56.19
Deep2Text-I[10, 11]	83.46	61.4	70.75	**83.46**	61.4	70.75	**83.46**	61.4	70.75
PAL[17, 16]	65.22	61.54	63.33	-	-	-	-	-	-
NJU	60.12	41.31	48.97	-	-	-	-	-	-
Baseline (OpenCv + Tessaract)[22]	46.48	37.13	41.28	47.2	32.82	38.72	30.29	24.2	26.9
TextCatcher-2	32.11	40.26	35.73	-	-	-	32.11	40.26	35.73
TextCatcher-1	11.58	5.01	6.99	-	-	-	11.58	5.01	6.99

The evaluation framework is similar to Task 3.1, but in this case an estimated word is considered a true positive if its intersection over union with a ground-truth word is larger than 0.5, and the word recognition is correct. Word recognition evaluation is case-insensitive. One- and two-character words are treated as *do not care*. Words containing non-alphanumeric characters are not taken into account with the exceptions of the hyphen and apostrophe. The recognition of punctuation marks at the beginning or the end of a ground truth word is optional and does not affect the evaluation.

The baseline offered is based on the TextSpotter framework for frame-by-frame detection (see section III-C), combined with the FoT tracker[8] of Tomas Vojir et al [25].

Five methods were submitted to this task. As can be seen in Table VII, it turns out that the "TextSpotter" baseline yields the highest ATA score. The winner method, "Stradvision", has very close numbers in MOTA and MOTP measures, indicating a similar performance in terms of detection, but the more than 10% lower ATA demonstrates a notable handicap in their tracking capabilities.

V. CHALLENGES 1 (BORN-DIGITAL IMAGES) AND 2 (FOCUSED SCENE TEXT)

Challenge 1 focuses on the extraction of textual content from born-digital images, while Challenge 2 addresses the scenario of focused text, which refers to images of text that has been explicitly focused on by the user. For details please refer to previous competition reports [4, 5, 6]. Tasks assessing End-to-End system performance were introduced for the 2015 edition. The task requires that all words in the image are both localised and recognised correctly.

Ground truth is defined at the word-level. Bounding boxes are axis-aligned rectilinear rectangles . One- or two-character words as well as words deemed unreadable are annotated in the dataset as *do not care*. Vocabularies were produced in the same manner as in Challenge 4. Similarly to Task 4.4 described before, three variants for the End-to-End task were defined, according to which vocabulary is provided during test time. The performance evaluation protocol is the same as in Task 4.4 (see Section III-C).

A. *TASK 1.4: End-to-End Systems for Born-Digital Images*

Eight methods were submitted to this task, the results of which are shown in Table VIII. Method ranking is based on *F-score* value. All metrics are calculated cumulatively over the whole test set. As a baseline method we use the "Baseline (OpenCV + Tesseract)" (see Section III-C).

[8] http://cmp.felk.cvut.cz/~vojirtom/

TABLE IX. RANKING IN TASK 2.4 (FOCUSED TEXT END-TO-END)

Method	Strong			Weak			Generic		
	P(%)	R(%)	F	P(%)	R(%)	F	P(%)	R(%)	F
VGGMaxBBNet[12]	89.63	**82.99**	**86.18**	-	-	-	-	-	-
Stradvision-1	88.66	75.03	81.28	83.98	**73.72**	**78.51**	69.46	64.99	67.15
Baseline (TextSpotter)[21]	85.91	69.79	77.02	61.68	64.78	63.19	50.91	58.12	54.28
Deep2Text-II[10, 11]	81.74	69.79	75.29	81.74	69.79	75.29	81.74	**69.79**	**75.29**
NJU	80.15	69.57	74.49	-	-	-	-	-	-
Deep2Text-I[10, 11]	83.95	66.74	74.36	83.95	66.74	74.36	**83.95**	66.74	74.36
MSER-MRF[26]	84.53	61.4	71.13	-	-	-	-	-	-
Beam Search CUNI	68.05	59.0	63.2	65.22	57.47	61.1	59.58	52.89	56.04
Baseline (OpenCv + Tessaract)[22]	75.72	48.96	59.47	69.45	47.11	56.14	50.96	37.62	43.29
Beam Search CUNI +S	**92.76**	15.38	26.38	**89.13**	13.41	23.32	65.48	12.0	20.28

The best performing method is “Stradvision-2”, without yielding top performances in either Precision or Recall. This is counter-intuitive considering that this method is adding a post-filtering step to “Stradvision-1”, which yields top performance in Precision. The best Recall is obtained by ”Deep2Text-II”, which is based on “USTB_TexStar” (see performance details in the 2013 edition report [6]) with an extra diversification step, coupled with a CNN-based recogniser. The “Deep2Text” variants do not seem to make use of the provided vocabularies as their performance remains the same for all levels of vocabulary contextualisation.

B. TASK 2.4: End-to-End Systems for Focused Scene Text

Eight methods were submitted to this task, the results of which are shown in Table IX. In this Task, we make use of both “Baseline (OpenCV + Tesseract)” and “Baseline (TextSpotter)” (see section III-C).

The best performing method is “VGGMaxBBNet”, which yields top performance in Recall and the second-best Precision score. The method is based on object proposals for localisation and a CNN-based recogniser. The best Precision score is obtained by “Beam Search CUNI +S” at the cost of the lowest obtained Recall score, similarly to Task 4.4. The “Deep2Text” variants do not seem to make use of the provided vocabularies.

VI. CONCLUSIONS

This report gives an overview of the ICDAR 2015 Robust Reading Competition. Up to date results are provided at the Web portal of the competition. The increased participation to the competition, as well as the continuous use of the competition’s Web portal (more than 750 registered users and 2,000 private submissions over the past 1.5 years) demonstrate the interest of the research community.

Compared to previous editions, persistent improvements can be observed although there is still a significant margin for improvement. On the methodological side, certain trends can be observed. First, we note that all submitted methods employ an initial segmentation step and text detection is obtained by classifying connected components or their groupings. Almost all methods make use of the MSER segmentation algorithm. Regarding text recognition, we note that top performing methods make use of commercial OCRs. This is in agreement to recent research that demonstrates that a conventional shape-based OCR engine is able to produce competitive results when provided with a conveniently preprocessed image.

ACKNOWLEDGMENTS

This work was supported by the projects TIN2014-52072-P, RYC-2009-05031, MASELTOV (FP7) and JST CREST.

REFERENCES

[1] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “ICDAR 2003 robust reading competitions,” in *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, vol. 2, 2003, pp. 682–687.

[2] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto *et al.*, “ICDAR 2003 robust reading competitions: entries, results, and future directions,” *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 2-3, pp. 105–122, 2005.

[3] S. M. Lucas, “ICDAR 2005 text locating competition results,” in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005, pp. 80–84.

[4] 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 *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1485–1490.

[5] A. Shahab, F. Shafait, and A. Dengel, “ICDAR 2011 robust reading competition challenge 2: Reading text in scene images,” in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1491–1496.

[6] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez i Bigorda, S. Robles Mestre, J. Mas, D. Fernandez Mota, J. Almazàn Almazàn, and L. P. de las Heras, “ICDAR 2013 robust reading competition,” in *Document Analysis and Recognition (ICDAR), 2013 International Conference on*. IEEE, 2013.

[7] D. Karatzas, S. Robles, and L. Gomez, “An on-line platform for ground truthing and performance evaluation of text extraction systems,” in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*. IEEE, 2014, pp. 242–246.

[8] H. I. Koo and D. H. Kim, “Scene text detection via connected component clustering and nontext filtering,” *Image Processing, IEEE Transactions on*, vol. 22, no. 6, pp. 2296–2305, 2013.

[9] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “Exploiting the circulant structure of tracking-by-detection with kernels,” in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 702–715.

[10] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, “Robust text detection in natural scene images,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 5, pp. 970–983, 2014.

[11] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, “Multi-orientation scene text detection with adaptive clustering,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, 2015.

[12] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Reading text in the wild with convolutional neural networks,” *arXiv preprint arXiv:1412.1842*, 2014.

[13] Z. Ze-Yu, T. Shu, and Y. Xu-Cheng, “Multi-strategy tracking based text detection in scene videos,” in *Document Analysis and Recognition (ICDAR), 2015 12th International Conference on (Submited for review)*. IEEE, 2015.

[14] D. Kumar, M. Prasad, and A. Ramakrishnan, “Maps: midline analysis and propagation of segmentation,” in *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*. ACM, 2012, p. 15.

[15] D. Kumar, M. A. Prasad, and A. Ramakrishnan, “Nesp: Nonlinear enhancement and selection of plane for optimal segmentation and recognition of scene word images,” in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013, pp. 865 806–865 806.

[16] C.-L. Liu, M. Koga, and H. Fujisawa, “Lexicon-driven segmentation and recognition of handwritten character strings for japanese address reading,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 11, pp. 1425–1437, 2002.

[17] Q.-F. Wang, F. Yin, and C.-L. Liu, “Handwritten chinese text recognition by integrating multiple contexts,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 8, pp. 1469–1481, 2012.

[18] L. Neumann and J. Matas, “Real-time scene text localization and recognition,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3538–3545.

[19] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1457–1464.

[20] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2014.

[21] L. Neumann and J. Matas, “On combining multiple segmentations in scene text recognition,” in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 523–527.

[22] L. Gómez and D. Karatzas, “Scene text recognition: No country for old men?” in *Computer Vision-ACCV 2014 Workshops*. Springer, 2014, pp. 157–168.

[23] L. Gomez and D. Karatzas, “Multi-script text extraction from natural scenes,” in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 467–471.

[24] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The CLEAR MOT metrics,” *EURASIP Journal on Image and Video Processing*, vol. 2008, May 2008.

[25] T. Vojır and J. Matas, “The enhanced flock of trackers,” *Registration and Recognition in Images and Videos*, vol. 532, p. 113, 2014.

[26] X. Liu, “Natural scene character recognition using markov random field,” in *Document Analysis and Recognition (ICDAR), 2015 12th International Conference on (Submited for review)*. IEEE, 2015.