2017 14th IAPR International Conference on Document Analysis and Recognition

ICDAR2017 Robust Reading Challenge on Omnidirectional Video

Masakazu Iwamura*, Naoyuki Morimoto*, Keishi Tainaka*,
Dena Bazazian†, Lluis Gomez†, Dimosthenis Karatzas†

*Dept. Computer Science and Intelligent Systems, Osaka Prefecture University, Sakai, Japan;
masa@cs.osakafu-u.ac.jp, {morimoto, tainaka}@m.cs.osakafu-u.ac.jp
†Computer Vision Centre, Universitat Autonoma de Barcelona, Barcelona, Spain;
{dbazazian, lgomez, dimos}@cvc.uab.es

***Abstract*—Results of ICDAR 2017 Robust Reading Challenge on Omnidirectional Video are presented. This competition uses Downtown Osaka Scene Text (DOST) Dataset that was captured in Osaka, Japan with an omnidirectional camera. Hence, it consists of sequential images (videos) of different view angles. Regarding the sequential images as videos (video mode), two tasks of localisation and end-to-end recognition are prepared. Regarding them as a set of still images (still image mode), three tasks of localisation, cropped word recognition and end-to-end recognition are prepared. As the dataset has been captured in Japan, the dataset contains Japanese text but also include text consisting of alphanumeric characters (Latin text). Hence, a submitted result for each task is evaluated in three ways: using Japanese only ground truth (GT), using Latin only GT and using combined GTs of both. Finally, by the submission deadline, we have received two submissions in the text localisation task of the still image mode. We intend to continue the competition in the open mode. Expecting further submissions, in this report we provide baseline results in all the tasks in addition to the submissions from the community.**

I. Introduction

Robust Reading refers to the automatic interpretation of written communication in unconstrained settings such as born-digital and real scene images and videos. The series of ICDAR Robust Reading Competitions (RRCs) have addressed the need to quantify and track progress in this domain since 2003 [1]–[6]. In 2017, along with other three challenges utilizing the RRCs' web portal [7], we organized a competition using Downtown Osaka Scene Text (DOST) Dataset [8], which features *scene texts in the wild*, *omnidirectional video* and *multi-scipt text (Japanese and Latin)*.

The DOST dataset contains videos (sequential images) captured in shopping streets in downtown Osaka with an omnidirectional camera. The use of an omnidirectional camera contributes to excluding user's intention in capturing images. Sequential images contained in the dataset contribute to encouraging developing a new kind of text detection and recognition techniques that utilize temporal information. Another important feature of DOST dataset is that it contains non-Latin text. Since the images were captured in Japan, a lot of Japanese text is contained while it also contains adequate amount of Latin text. Because of these features, we say that the DOST dataset fits well the setting of scene texts in the wild scenario.

In the organisation, the release of the DOST dataset was released by June 18 and submission site was opened by June 27. The submission deadline was set on June 30. Hence, participants can spend only limited days for preparing and evaluating their methods. Regardless of the situation, by the submission deadline, we have received two submissions in the text localisation task of the still image mode. We intent to continue the competition in the open mode. Expecting further submissions, in this report we provide baseline results in all the tasks. The source code of the baseline method is made publicly available[1].

II. Dataset and Tasks

Table I shows the detail of the DOST dataset released for the competition. Video sequences are divided into four places. Data of places 1-3 are for testing and those of place 4 for training. Since the images are captured with an omnidirectional camera equipping six cameras, six images with different views are available for each place. Image sequences of some cameras are divided into multiple because only a part of the sequential images are ground truthed. As a result, the dataset consists of "sequences," each of which comprises consecutive images captured with a single camera. Ground truth (GT) of the dataset has been under renovation. As of the competition period, GT of only place 1 was renovated. The GT files will be updated when the renovation is completed. When GT is updated, already submitted results will be automatically re-evaluated on the new GT.

In addition to the DOST dataset, we provide images of Japanese characters in multiple fonts for training. Participants are allowed to use any training samples and requested to mention which samples are used for training in the submission.

Considering the nature of the DOST dataset, we prepared following two modes for the competition: video mode and still image mode.

[1]http://github.com/ComputerVisionCentre/RRC2017_DOST_Baseline

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.236

TABLE I
DETAILS OF DOST DATASET RELEASED.

Place	Train/Test	Camera		Sequence and #frame
1	Test	Side camera	0	Seq#1: 1501 frames
			1	Seq#1: 243 frames Seq#2: 47 frames Seq#3: 232 frames
			2	Seq#1: 582 frames Seq#2: 48 frames
			3	Seq#1: 263 frames Seq#2: 25 frames
			4	Seq#1: 303 frames Seq#2: 72 frames
		Top camera	5	N/A
2	Test	Side camera	0	Seq#1: 5001 frames
			1	Seq#1: 2000 frames
			2	Seq#1: 2009 frames
			3	Seq#1: 2001 frames
			4	Seq#1: 2004 frames
		Top camera	5	N/A
3	Test	Side camera	0	Seq#1: 5001 frames
			1	Seq#1: 1501 frames Seq#2: 8 frames Seq#3: 1 frame
			2	Seq#1: 1950 frames
			3	Seq#1: 2001 frames
			4	Seq#1: 2001 frames
		Top camera	5	No GT available
4	Train	Side camera	0	Seq#1: 3193 frames Seq#2: 32 frames Seq#3: 101 frames Seq#4: 29 frames
			1	N/A
			2	
			3	
			4	
		Top camera	5	

A. Video mode

In the video mode, we regard the DOST dataset as a video dataset (i.e., each image sequence is regarded as a single video) and treat it in the same manner as the "Text in Videos" Challenge of ICDAR 2013/2015 RRC [5], [6]. The dataset is provided in a similar way to the "Text in Videos" Challenge of ICDAR 2013/2015 RRC. The only difference is that the GT of the DOST dataset has a tag representing script (Latin or Japanese) instead of the tag representing language. For each time, participants are allowed to submit a single file which contains result of Latin only, Japanese only or both. The submitted result is automatically evaluated in three modalities (Latin only, Japanese only and both). We did not provide vocabularies specially prepared for the tasks.

In this mode, we organise the following two tasks:

1) Task V1: Localisation: The objective of this task is the correct localisation and tracking of words (excluding "do not care" ones) in the sequence.
2) Task V2: End-to-end: This task aims to assess End-to-End system performance that combines correct localisation and tracking with correct recognition.

B. Still image mode

In the still image mode, we regard the DOST dataset as a set of still images and treat it in a similar manner as the "Incidental Scene Text" Challenge of ICDAR 2015 RRC [6]. Again, the only difference is that the GT of the DOST dataset has a tag representing script (Latin or Japanese). In the localisation task (Task I1) and the end-to-end task (Task I3), datasets consist of frame images sampled every 10 frames from the video sequences are provided. In the cropped word recognition task (Task I2), selected cropped images are provided. We did not provide vocabularies specially prepared for the tasks.

In this mode, we organise the following three tasks:

1) Task I1: Localisation: The objective of this task is the correct localisation of words (excluding "do not care" ones) of the image.
2) Task I2: Cropped word recognition: This task aims to evaluate recognition performance over a set of pre-localised word regions.
3) Task I3: End-to-end: This task aims to assess End-to-End system performance that combines correct localisation and correct recognition.

III. RESULTS OF STILL IMAGE MODE

A. Task I1: Localisation in still image

Text localisation task in still images has attracted much attention. In evaluation, we basically followed the "Incidental Scene Text" Challenge of ICDAR 2015 RRC [6].

In this task, two results are submitted. From the description of the methods, SCUT-DLVClab is a CNN-based method submitted by Yuliang Liu, Sheng Zhang and Lianwen Jin. NLPR-PAL submitted by Wenhao He, Fei Yin, Cheng-Lin Liu is based on [9]. They also use images of Chinese, Japanese and Korean from the ICDAR2017 Competition on Multi-lingual scene text detection and script identification.

In addition, we prepared two methods as baseline methods. One is Optical character recognition (OCR) of Google Cloud Platform[2]. The other is the OpenCV[3] implementation of the text detection algorithm proposed by Neumann and Matas [10] combined with the Tesseract 3.0 OCR engine[4]. The OpenCV module is used for extracting textline bounding box and the Tesseract is used for dividing the textline bounding box into word bounding box and filtering out false positives.

Table II shows the results. Overall, looking at the results of each single image, all methods faced difficulty in finding text. In other words, achieving a high recall rate was more difficult than achieving a high precision rate. Typical errors which are caused by two features of Japanese text were (1) a word region is misplaced and (2) a single word region is detected as multiple word regions. As shown in Fig. 1, the former is caused by the fact that often a space or a clear boundary does not exist between words or alternative units we used for Japanese text (bunsetsu[5]). This observation implies that it would be difficult to solve this problem without a vocabulary.

[2]https://cloud.google.com/vision/docs/ocr

[3]http://opencv.org

[4]http://github.com/tesseract-ocr

[5]Bunsetsu is the smallest unit of words that sounds natural in a spoken sentence.

TABLE II
RESULTS OF (TASK I1) LOCALISATION IN STILL IMAGE MODE, EXPRESSED IN PERCENT. BOLD TEXT INDICATES THE BEST. "*" INDICATES A BASELINE METHOD.

Evaluation	Method	Recall	Precision	F-score
Global	SCUT-DLVClab	11.60	**46.41**	18.56
	NLPR-PAL	**16.25**	38.10	**22.78**
	Google Cloud Platform*	4.09	14.67	6.40
	OpenCV + Tesseract*	1.55	7.95	2.60
Latin	SCUT-DLVClab	12.26	**28.37**	**17.12**
	NLPR-PAL	**13.31**	17.31	15.05
	Google Cloud Platform*	5.61	13.92	8.00
	OpenCV + Tesseract*	2.92	5.20	3.74
Japanese	SCUT-DLVClab	11.31	**39.79**	17.62
	NLPR-PAL	**17.49**	32.66	**22.78**
	Google Cloud Platform*	3.44	10.70	5.21
	OpenCV + Tesseract*	0.97	3.78	1.54

Fig. 1. Ground truths in a region.

Fig. 2. Detection result of NLPR-PAL for the region of Fig. 1.

For example, Fig. 2 is the detection result of NLPR-PAL for the region of Fig. 1, in which the detected region covers a part of two word regions connected without a clear boundary[6]. The latter is caused by relatively long distance between characters forming a word.

Comparing submitted and baseline methods, submitted results were far better than the baseline results. Comparing the two submissions, SCUT-DLVClab tended to be better in precision and NLPR-PAL better in recall in all three evaluation modalities. Regarding F-score, NLPR-PAL achieved better in the Japanese evaluation and SCUT-DLVClab better in the Latin evaluation. It seems that since the number of Japanese words is larger than that of Latin words, NLPR-PAL achieved better in the global evaluation (mixture of Japanese and Latin). According to [9], NLPR-PAL achieved the F-score of 81% in "Incidental Scene Text" Challenge of ICDAR 2015 RRC and is the state-of-the-art. Comparing with the figure, figures achieved in this task were much smaller, which indicates that this task is more difficult than the "Incidental Scene Text" dataset.

Results of Google Cloud Platform were not as good as we expected. Looking at the results of each single image, it tended

[6]Note that other methods could not extract any word region.

Fig. 3. Examples of vertically aligned text.

to achieve a high precision with low recall. While we do not have an access to the detail of the method, a possible reason is that the method is designed for text recognition and word candidates with low confidences (low recognition scores) are suppressed. Similarly, the baseline method *OpenCV + Tesseract* performs poorly in this task, while it is known to achieve much better in results in the "Incidental Scene Text" dataset [6], [11]. This shows the limitations of text detection and recognition algorithms that are designed for English-only and well-focussed horizontal text in mind when applied to a less constrained scenario like the one of the DOST dataset.

B. Task I2: Cropped word recognition in still image

Cropped word recognition task is often performed with a vocabulary provided. In such a case, the task is to select the most feasible word from the word list and called word spotting. In this task, since we did not provide vocabulary, it is more like *pure* word recognition task, while participants still can use a vocabulary if they prepare it by themselves. In this task, we manually selected target words from the DOST dataset. Word images selected from places 1 and 2 are used for testing (the number of words were 320 in Japanese and 165 in Latin) and ones from place 4 are for training (69 in Japanese and 61 in Latin). In evaluation, we basically followed the "Incidental Scene Text" Challenge of ICDAR 2015 RRC [6].

Since no submission was made for this task till the submission deadline for this task, we show the results of three baseline methods: The OCR of Google Cloud Platform (used also in the localisation task), the Tesseract OCR engine version 4.00alpha with the new Neural Nets (LSTM) engine mode, and the DictNet CNN model of Jaderberg *et al.* [12]. In all cases cropped word images are sent to the recognition engine without any further processing. Note that the DictNet is an English-only dictionary based word spotting model. Hence, it can read English text only.

Table III shows the results. Overall, DictNet was the best in the Latin evaluation and Google Cloud Platform was the best in the Japanese evaluation. As the number of Japanese words is larger than that of Latin words, in the global evaluation, Google Cloud Platform achieved the best performance. Checking the results of individual methods (Table IV shows selected individual results), Google Cloud Platform did not output any transcription in 355 images (73% of 485 images; 242 images in Japanese and 113 images in Latin). As observed in the localisation task, this also indicates that the method is tuned

TABLE III
RESULTS OF (TASK I2) CROPPED WORD RECOGNITION IN STILL IMAGE MODE. "*" INDICATES A BASELINE METHOD.

Evaluation	Method	Total Edit distance	Correctly Recognised Words (%)
Global	Google Cloud Platform*	**396.36**	**11.13**
	DictNet [12]*	477.93	8.45
	Tesseract OCR 4.00*	459.10	7.42
Latin	Google Cloud Platform*	121.81	18.79
	DictNet [12]*	**103.22**	**24.24**
	Tesseract OCR 4.00*	132.36	18.79
Japanese	Google Cloud Platform*	**274.55**	**7.19**
	DictNet [12]*	374.71	0.31
	Tesseract OCR 4.00*	326.75	1.56

not to output less reliable transcriptions. Tesseract and DictNet looked they cannot segment characters of vertically aligned text (examples are shown in Fig. 3) because the number of characters of the outputted transcription for vertically aligned texts were typically one or two characters and wrong. While DictNet clearly outperformed the other methods in the Latin evaluation, the percentage of correctly recognised words is less than 25%. Thus, there is much room to improve. Note that despite DictNet can recognise only English text, its "Correctly Recognised Words" was not zero percent. This was caused by mis-categorization of Latin-only words into Japanese text.

C. Task I3: End-to-end system in still image

This task aims to evaluate end-to-end system performance in the setting of scene texts in the wild scenario. The evaluation strategy combines measuring localisation efficiency and recognition capacity over all care words. In evaluation, we basically followed the "Incidental Scene Text" Challenge of ICDAR 2015 RRC [6].

Since no submission was made for this task till the submission deadline, we show only the results of two baseline methods used also in the localisation task: the OCR of Google Cloud Platform, and the scene OpenCV scene text detection module combined with the Tesseract OCR engine.

Table V shows the results. As the same baseline methods as the localisation task were used, the obtained results had the same tendency. That is, Google Cloud Platform achieved better performance than OpenCV + Tesseract. Compared with the top result of the "Incidental Scene Text" Challenge of ICDAR 2015 RRC [6] (i.e., 43.7%), achieved performance in this task was far lower; the best F-scores achieved by Google Cloud Platform were 3.81%, 5.56% and 2.84% in the global, Latin and Japanese evaluations, respectively. This indicates that there is very big room to improve.

IV. RESULTS OF VIDEO MODE

A. Task V1: Localisation in video

This task requires that words are both detected and tracked correctly over the video sequence. Following "Text in Video" Challenge of ICDAR 2013/2015 RRCs, the evaluation is based on an adaptation of the CLEAR-MOT framework [13] 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 ICDAR 2013 RRC report [5] for details about these metrics.

Since no submission was made for this task till the deadline, we show only the results of the baseline method that is similar to the one used in "Text in Video" Challenge of ICDAR 2013/2015 RRCs [5], [6]. The baseline algorithm consists of three different stages: text detection, tracking, and recognition. The detection stage is performed with the OpenCV implementation of the text detection algorithm proposed by Neumann and Matas [10]. The tracking stage is performed with the OpenCV implementation of the KCF tracking algorithm [14] with color-names features [15]. For the text recognition stage, the baseline method makes use of the Tesseract 3.0 OCR engine.

Table VI shows the results. As shown in the table, the mean tracking precision (MOTP) of the baseline algorithm was 53.94, and the mean tracking accuracy (MOTA) was -51.14. Notice that negative values of MOTA were obtained when the evaluated method counts more false positives and/or ID-switches per frame than the actual number of words in the ground-truth. The low MOTA and ATA values obtained were consequence of a large number of false positives and a high fragmentation in object's IDs.

The obtained results show coherency with the ones reported for a similar shallow baseline on the ICDAR 2013 *Text in Videos* challenge [5] while being lower in general. This indicates that the the baseline method is not well suited for the challenging scenario of the DOST dataset.

B. Task V2: End-to-end system in video

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 evaluation framework is similar to Task V1, 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.

Since no submission was made for this task till the deadline, we prepared a baseline method. However, we decided to omit it because the result was not reliable enough.

V. CONCLUSION

This report gives an overview of the ICDAR2017 Robust Reading Challenge on Omnidirectional Video. Due to serious

TABLE IV
SELECTED RESULTS OF (TASK I2) CROPPED WORD RECOGNITION IN STILL IMAGE MODE.

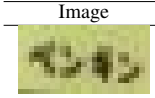
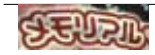
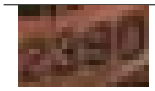
Image	Ground truth	Google Cloud Platform	DictNet [12]	Tesseract OCR 4.00
	ペンギン		REAP	ペンキ#ン
	基本料金	基本	KIT	[もす-
	全身マッサージ		ESPYING	Rb
	メモリアル		RUM	モUつ
	まいどおおきに		ASTER	さ絵さ己
	高価		FACTOID	高側
	ぱそまる		BIFFS	ばとまきる
	ボークス	オークス	Z	ー享
	とんかつ	とんかつ	LINING	ぞどんあふっ
	年中無休	年中	STOPOVER	はcね体
	300種類		BODKIN	800稀然
	ROCKY		ROCKY	oc-
	JEWELRY	EMELBY	EEG	pe?
	DAIKOKUYA		DROWN	DAKOWUA
	WANTED		MEWLED	|\!\\\|Hl|
	SALE		SALE	22m
	SLOT		S	C
	Rest		RESP	Say
	NORTHERN		REPETITIVENESS	E cl
	0570		OSLO	NK
	2390		EGGO	>390

TABLE V
RESULTS OF (TASK I3) TASK I3: END-TO-END SYSTEM IN STILL IMAGE MODE, EXPRESSED IN PERCENT. "*" INDICATES A BASELINE METHOD.

Evaluation	Method	Recall	Precision	F-score
Global	Google Cloud Platform*	**2.35**	**10.05**	**3.81**
	OpenCV + Tesseract*	0.25	1.29	0.42
Latin	Google Cloud Platform*	**3.56**	**12.63**	**5.56**
	OpenCV + Tesseract*	0.82	1.46	1.05
Japanese	Google Cloud Platform*	**1.83**	**6.33**	**2.84**
	OpenCV + Tesseract*	0.01	0.04	0.01

TABLE VI
RESULTS OF (TASK V1) LOCALISATION IN VIDEO MODE, EXPRESSED IN PERCENT. "*" INDICATES A BASELINE METHOD.

Evaluation	Method	MOTP	MOTA	ATA
Global	OpenCV + Tesseract*	53.94	-51.14	0.18
Latin	OpenCV + Tesseract*	54.16	-61.8	0.19
Japanese	OpenCV + Tesseract*	52.97	-64.2	0.16

delay of releasing the dataset, the total number of submissions made for the competition were only two. However, we plan to reopen the competition site in the open mode and keep providing up to date results at the Web portal of the competition. We hope this article and the DOST dataset contribute to promote improvement of robust reading techniques.

Acknowledge

The authors appreciate Prof. Koichi Kise of Osaka Prefecture University for his various support. This work is supported by Spanish project TIN2014-52072-P, the CERCA Programme / Generalitat de Catalunya, JST CREST and JSPS KAKENHI #17H01803 and #25240028.

REFERENCES

[1] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. ICDAR*, vol. 2, Aug. 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, H. Miyao, J. Zhu, W. Ou, C. Wolf, J.-M. Jolion, L. Todoran, M. Worring, and X. Lin, "ICDAR 2003 robust reading competitions: Entries, results and future directions," *IJDAR*, vol. 7, no. 2-3, pp. 105–122, 2005.

[3] S. M. Lucas, "ICDAR 2005 text locating competition results," in *Proc. ICDAR*, vol. 1, Aug. 2005, pp. 80–84.

[4] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Proc. ICDAR2011*, 2011, pp. 1491–1496.

[5] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *Proc. ICDAR*, 2013, pp. 1115–1124.

[6] 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 robust reading competition," in *Proc. ICDAR*, 2015, pp. 1156–1160.

[7] D. Karatzas, S. Robles, and L. Gomez, "An on-line platform for ground truthing and performance evaluation of text extraction system," in *Proc. DAS*, Apr. 2014, pp. 242–246.

[8] M. Iwamura, T. Matsuda, N. Morimoto, H. Sato, Y. Ikeda, and K. Kise, "Downtown osaka scene text dataset," in *Proc. IWRR2016*, ser. Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016, pp. 440–455.

[9] F. Y. Wenhao He, Xu-Yao Zhang and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," *arXiv:1703.08289 [cs.CV]*, 2017. [Online]. Available: https://arxiv.org/abs/1703.08289

[10] 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.

[11] L. Gómez and D. Karatzas, "Scene text recognition: No country for old men?" in *Asian Conference on Computer Vision*. Springer, 2014, pp. 157–168.

[12] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in *Workshop on Deep Learning, NIPS*, 2014.

[13] 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.

[14] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[15] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1090–1097.