Downtown Osaka Scene Text Dataset

Masakazu Iwamura(✉), Takahiro Matsuda, Naoyuki Morimoto, Hitomi Sato,
Yuki Ikeda, and Koichi Kise

Department of Computer Science and Intelligent Systems,
Graduate School of Engineering, Osaka Prefecture University, Sakai, Japan
{masa,kise}@cs.osakafu-u.ac.jp, morimoto@m.cs.osakafu-u.ac.jp

Abstract. This paper presents a new scene text dataset named Downtown Osaka Scene Text Dataset (in short, DOST dataset). The dataset consists of sequential images captured in shopping streets in downtown Osaka with an omnidirectional camera. Unlike most of existing datasets consisting of scene images intentionally captured, DOST dataset consists of uncontrolled scene images; use of an omnidirectional camera enabled us to capture videos (sequential images) of whole scenes surrounding the camera. Since the dataset preserved the real scenes containing texts as they were, in other words, they are *scene texts in the wild*. DOST dataset contained 32,147 manually ground truthed sequential images. They contained 935,601 text regions consisting of 797,919 legible and 137,682 illegible. The legible regions contained 2,808,340 characters. The dataset is evaluated using two existing scene text detection methods and one powerful commercial end-to-end scene text recognition method to know the difficulty and quality in comparison with existing datasets.

Keywords: Scene text in the wild · Uncontrolled scene text · Omnidirectional camera · Sequential image · Video · Japanese text

1 Introduction

Text plays important roles in our life. Imagining life in a world without text, in which, for example, neither book, newspaper, signboard, menu in a restaurant, texting on smartphone nor program source code exists or they exist in a completely different form, we can rediscovery not only the necessity of text but also importance of reading and interpreting text. Although only human being has been endowed with the ability of reading and interpreting text, researchers have struggled to enable computers to read text.

Focusing on camera-captured text and scene text, some pioneer works were presented in 1990s [21]. Since then, increasing attention was paid for recognizing scene text. Table 1 shows remarkable recent progress of scene text recognition techniques. In the table, most of reported accuracies of the latest methods exceeded 90 % on major benchmark datasets. However, does this mean these methods are powerful enough to read a variety of texts in the real environment? Many people would agree that the answer is no. Text images contained in these

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part I, LNCS 9913, pp. 440–455, 2016.
DOI: 10.1007/978-3-319-46604-0_32

Table 1. Recent improvement of recognition performance in scene text recognition tasks. Based on Table 1 of [1], this table summarizes recognition accuracies of recent methods in percentage terms on representative benchmark datasets in the chronological order. "50," "1k" and "50k" represent lexicon sizes. "Full" and "None" represent with all per-image lexicon words and without lexicon, respectively.

Year	Method	IIIT5K [2]			SVT [3]		ICDAR03 [4]				ICDAR13 [5]
	Lexicon	50	1k	None	50	None	50	Full	50k	None	None
-	ABBYY [3]	24.3	-	-	35.0	-	56.0	55.0	-	-	-
2011	Wang et al. [3]	-	-	-	57.0	-	76.0	62.0	-	-	-
2012	Mishra et al. [2]	64.1	57.5	-	73.2	-	81.8	67.8	-	-	-
	Wang et al. [6]	-	-	-	70.0	-	90.0	84.0	-	-	-
	Novikova et al. [7]	-	-	-	72.9	-	-	82.8	-	-	-
2013	Goel et al. [8]	-	-	-	77.3	-	89.7	-	-	-	-
	Bissacco et al. [9]	-	-	-	90.4	78.0	-	-	-	-	87.6
2014	Alsharif and Pineau [10]	-	-	-	74.3	-	93.1	88.6	85.1	-	-
	Almazán et al. [11]	91.2	82.1	-	89.2	-	-	-	-	-	-
	Yao et al. [12]	80.2	69.3	-	75.9	-	88.5	80.3	-	-	-
	Jaderberg et al. [13]	-	-	-	86.1	-	96.2	91.5	-	-	-
	Su and Lu [14]	-	-	-	83.0	-	92.0	82.0	-	-	-
2015	Rodrguez-Serrano et al. [15]	76.1	57.4	-	70.0	-	-	-	-	-	-
	Gordo [16]	93.3	86.6	-	91.8	-	-	-	-	-	-
	Jaderberg et al. [17]	97.1	92.7	-	95.4	80.7	98.7	98.6	93.3	93.1	90.8
	Jaderberg et al. [18]	95.5	89.6	-	93.2	71.7	97.8	97.0	93.4	89.6	81.8
	Shi et al. [19]	97.6	94.4	78.2	96.4	80.8	98.7	97.6	95.5	89.4	86.7
2016	Shi et al. [1]	96.2	93.8	81.9	95.5	81.9	98.3	96.2	94.8	90.1	88.6
	Poznanski and Wolf [20]	97.9	94.2	-	96.6	83.6	-	-	-	-	-

datasets are far easier than the real. In the real environment, scene text is more diverse; for example, various designs/styles/shapes of texts under many different illuminations are taken from variety of angles/distances. In this regard, there is a big gap between scene texts contained in these existing datasets and observed in the real environment.

In this paper, to fill the gap, we present a new dataset named Downtown Osaka Scene Text Dataset (in short, DOST dataset) that preserved scene texts observed in the real environment as they were. The dataset contains videos (sequential images) captured in shopping streets in downtown Osaka with an omnidirectional camera equipped with five horizontal and one upward cameras shown in Fig. 1. In total, 30 image sequences (consisting of five shopping streets times six cameras) consisting of 783,150 images were captured. Among them, 27 image sequences consisting of 32,147 images were manually ground truthed. As a result, 935,601 text regions consisting of 797,919 legible and 137,682 illegible text regions were obtained. The legible regions contained 2,808,340 characters. Since the images were captured in Japan, they contained many Japanese texts. However, out of the whole (797,919) legible text regions, 283,940 consisted of only alphabets and digits. These legible text regions contained 1,138,091 non-Japanese characters. Because of the above mentioned features of the dataset, we can say that DOST dataset preserved *scene texts in the wild*. Figures 3, 4, 5 and 6 show examples of captured images ground truthed and segmented words

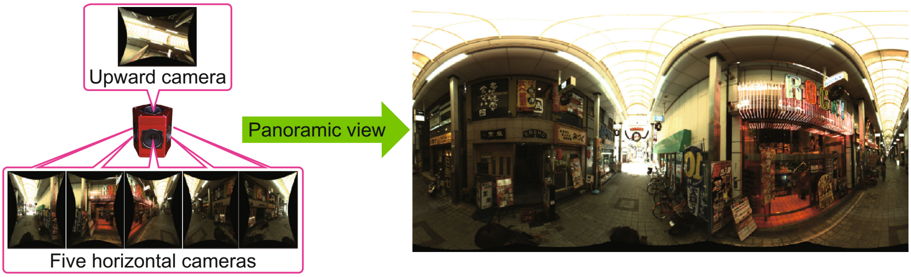


Fig. 1. Point Grey Ladybug3, an omnidirectional camera, captures six images consisting of five horizontal and one upward cameras at once. A panoramic view can be created from the six images.

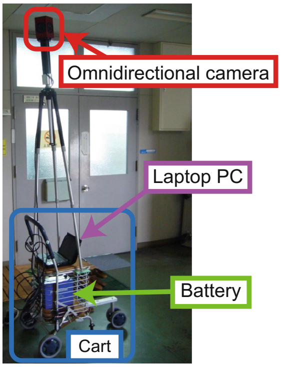


Fig. 2. Equipment used for capturing.

contained in DOST dataset. Since the sequence images were captured with an omnidirectional camera and continuous in time, a single word was captured many times in multiple view angles. The DOST dataset was evaluated using two existing text detection and one powerful commercial end-to-end scene text recognition methods to measure the difficulty and quality in comparison with existing datasets.

2 Unique Features of DOST Dataset

Features of existing datasets are summarized in Table 2. Major differences of DOST dataset from existing datasets include following.

1. DOST dataset contains only real images. Unlike MJSynth [22] and SynthText [23] aiming at training a better classifier, DOST dataset aims at evaluation of scene text detection/recognition methods.
2. The images were completely not intentionally captured. In this regard, the most similar dataset is the one dedicated to ICDAR 2015 Robust Reading

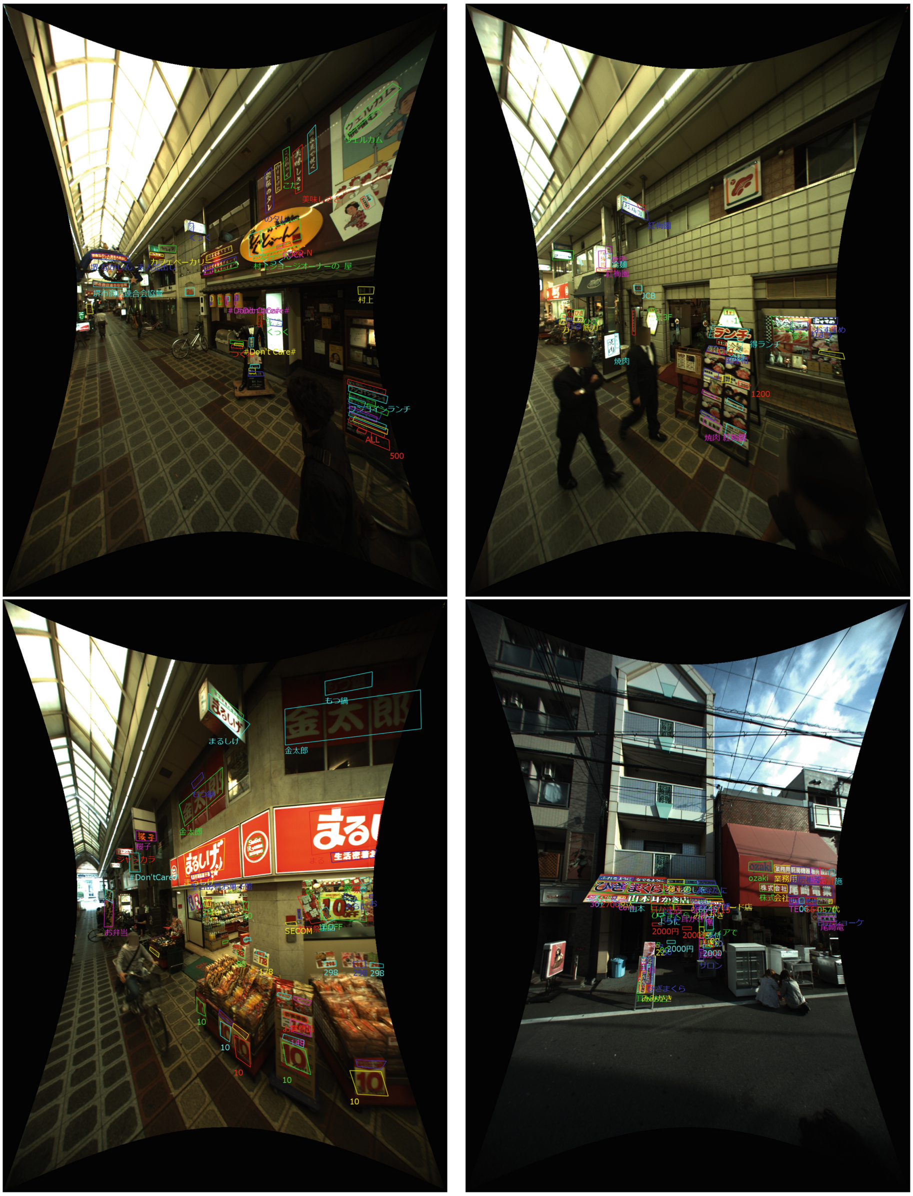


Fig. 3. Samples of captured images ground truthed. The four images in this page are selected from ones ground truthed. Bounding boxes represent word regions and texts next to bounding boxes text annotations.

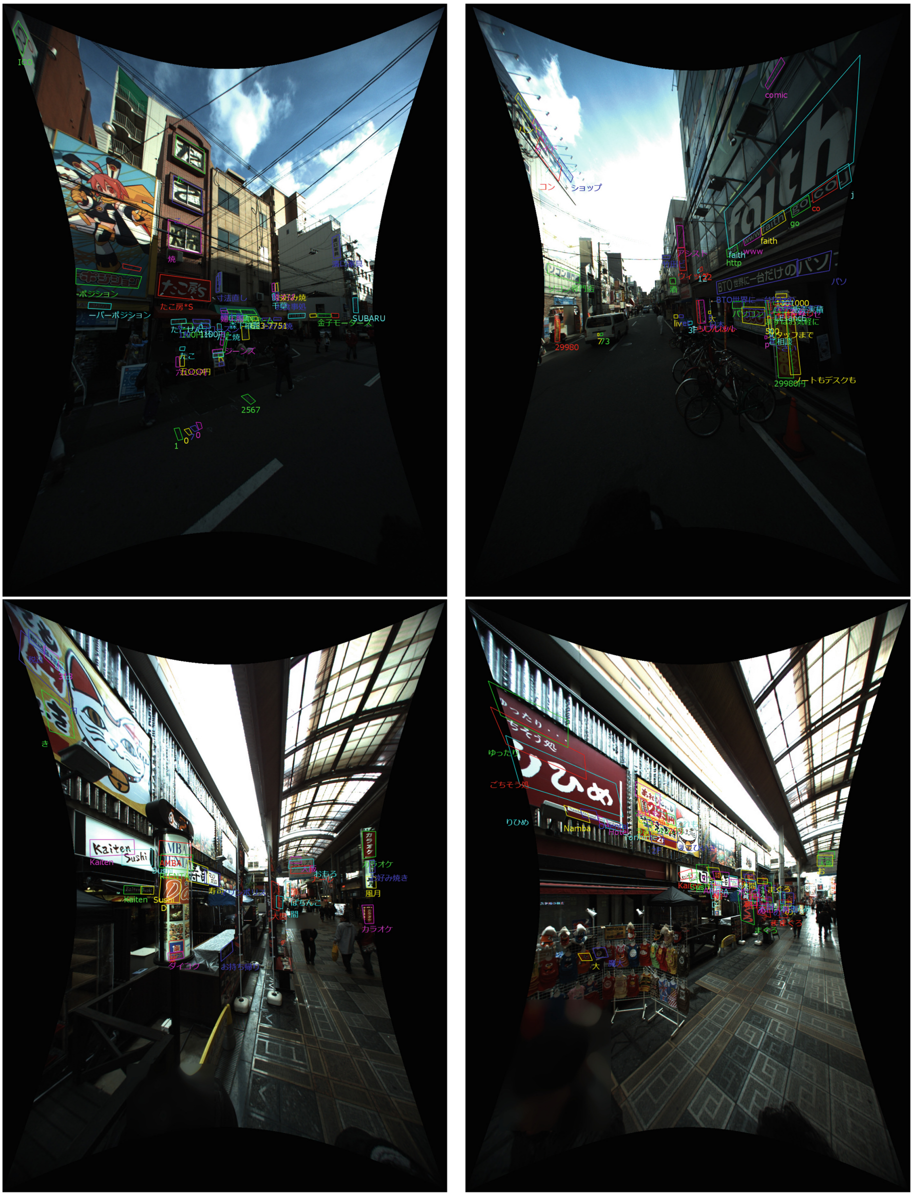


Fig. 4. Samples of captured images ground truthed (continued). The four images in this page are selected from ones ground truthed. Bounding boxes represent word regions and texts next to bounding boxes text annotations.

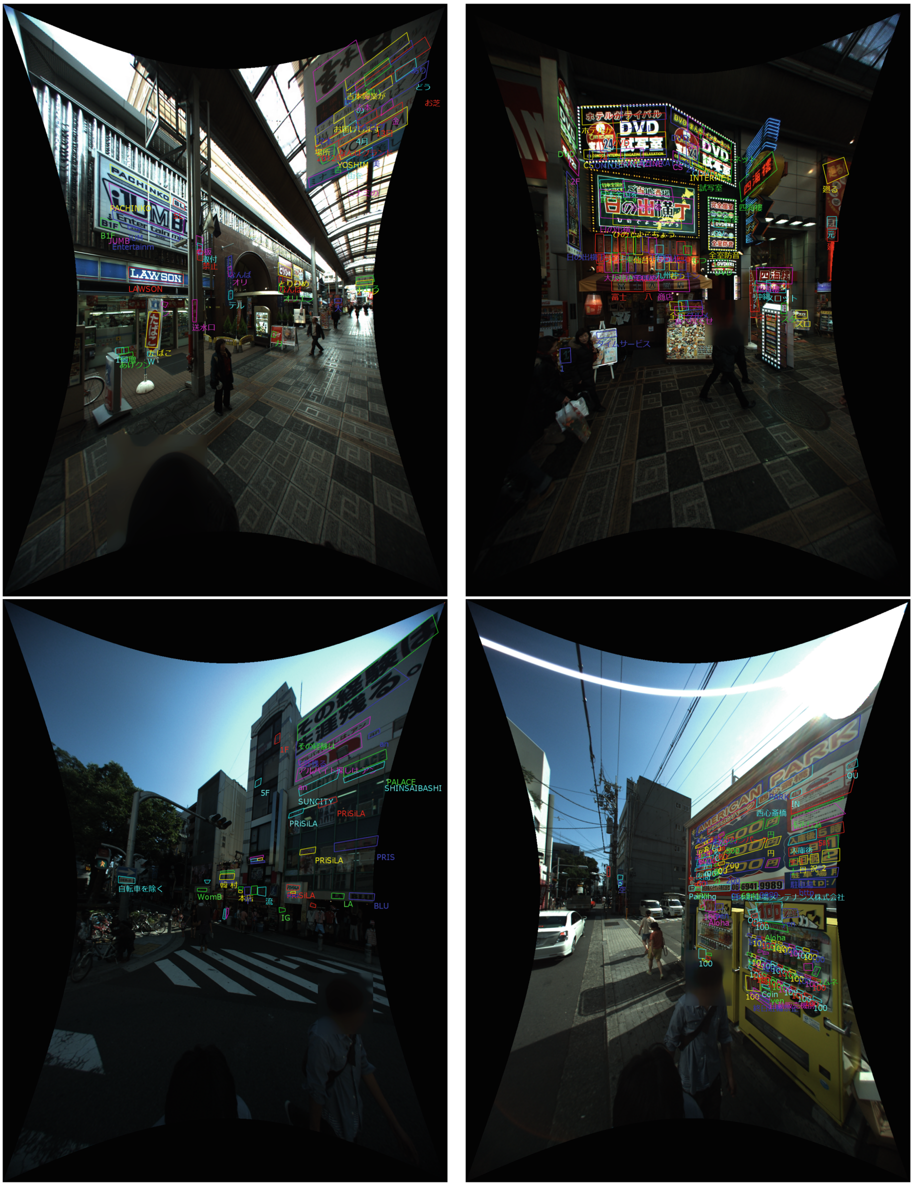


Fig. 5. Samples of captured images ground truthed (continued). The four images in this page are selected from ones ground truthed. Bounding boxes represent word regions and texts next to bounding boxes text annotations.

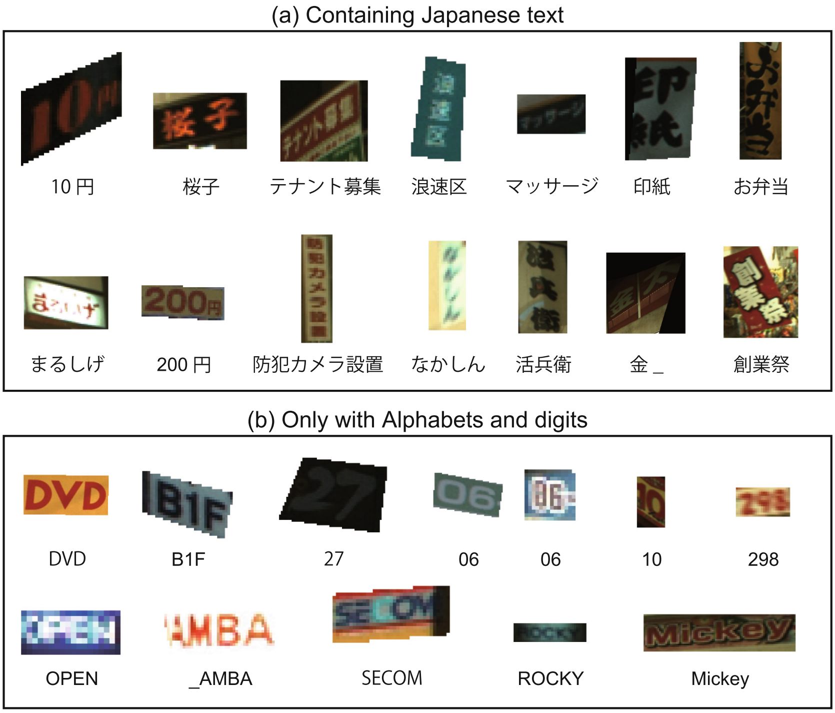


Fig. 6. Samples of segmented words contained in DOST dataset. "_" means there is partially occluded character(s).

Competition Challenge 4 "incidental scene text." It is regarded not intentionally captured because images in the dataset were captured with Google Glass without having taken any prior action to cause its appearance in the field of view or improve its positioning or quality in the frame. DOST dataset is completely free from intention even from face direction of the user wearing Google glass.

3. The images are a video dataset (consecutive in time). There are already video datasets. The 2013 and 2015 editions of ICDAR Robust Reading Competition (RRC) Challenge 3 datasets [5,24] consists of sequential images. The biggest difference is that DOST dataset was captured with an omnidirectional camera. Another difference is that DOST dataset contains Japanse text while ICDAR RRC datasets consists of Latin text. Another video dataset YVT [25] contained YouTube videos. Some texts in the dataset are not scene texts but just captions.
4. DOST dataset contains multiple word images of a single word taken in different view angles.
5. The scale of DOST dataset is large. In the following discussion, let us exclude synthesized datasets and SVHN consisting of digit. Though the number of total images ground truthed in DOST dataset (32,147) is not very large

(almost half of the largest dataset, COCO-Text), the number of word regions (935,601 in total consisting of 797,919 legible and 137,682 illegible) is very large (a factor of 4.6 times larger than the second largest dataset, COCO-Text). This is because image sizes are relatively large ($1,200 \times 1,600$ pixels) and the images were captured in shopping streets where a lot of texts exist. DOST dataset is also the largest in terms of the number of unique word sequences, which is larger than the second largest, ICDAR2015 Challenge 3 dataset, by a factor of 6.3 times.

Another feature of DOST dataset is that it was manually ground truthed by students. The reason we did not use a crowdsourcing service such as Amazon Mechanical Turk[1] is most of workers cannot read Japanese text.

Yet another feature of DOST dataset is that it contains many Latin characters, though the images were captured in Japan. The number of characters per category and examples of Japanese characters and symbols are shown in Fig. 7. Kanji (aka. Chinese character) is a logogram. Katakana and Hiragana are syllabaries invented based on Kanji. Though symbols are originally not intended to be ground truthed, some were actually ground truthed. They include often used iteration marks such as “‘々’” which represent a duplicated character. In the future, other than the iteration marks would be discarded by rigorously applying the ground truthing policy.

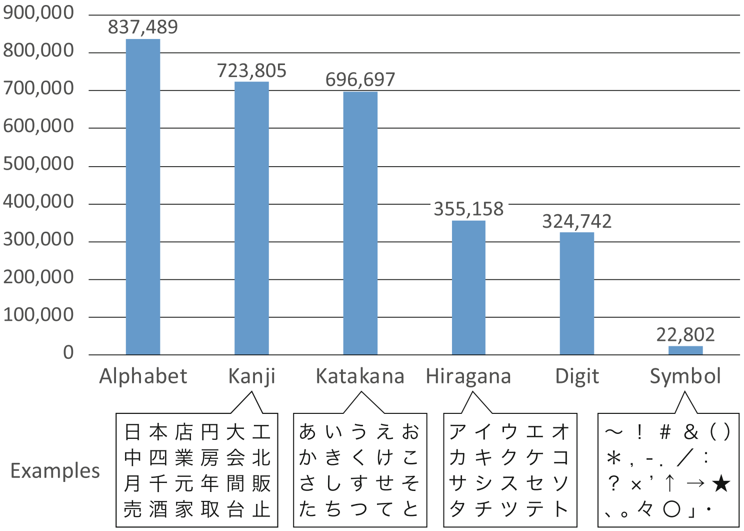


Fig. 7. Number of characters per category and examples of Japanese characters and symbols.

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

Table 2. Summary of publicly available datasets. "Video?" is whether the images are consecutive in time. "Real?" is whether the dataset consists of real images only (Yes) or not (No; note that captions are regarded as synthesized). #Image represents the total number of images (for a video dataset, the total number of frames). #Word represents the number of word regions ground truthed. In a video dataset, #WS represents the number of word sequences which do not consist of only "don't care" regions.

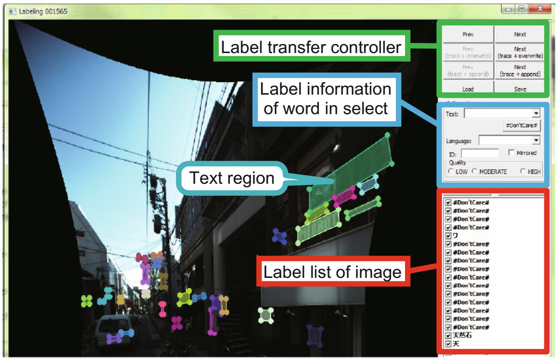
Name	Real?	Video?	#Image	#Word	#WS	Language
ICDAR2003 [4]	Yes	No	509	2,268	-	English
ICDAR2013 (Challenge 2)[5]	Yes	No	462	2,524	-	English
ICDAR2013 (Challenge 3) [5]	Yes	Yes	15,277	93,598	1,962	English, French, Spanish
ICDAR2015 (Challenge 3) [24]	Yes	Yes	27,824	125,141	3,562	English, French, Spanish
ICDAR2015 (Challenge 4) [24]	Yes	No	1,670	17,548	-	English
NEOCR [26]	Yes	No	659	5,238	-	English, German
KAIST [27]	Yes	No	3,000	3,000	-	English, Korean
SVT [3]	Yes	No	349	904	-	English
SVHN [28]	Yes	No	248,823	630,420[a]	-	Digit
IIIT5K [2]	Yes	No	5,000	5,000	-	English
YVT [25]	No	Yes	11791	16,620	245	English
MJSynth [22]	No	No	8,919,273	8,919,273	-	English
COCO-Text [29]	Yes	No	63,686	173,589	-	English, Germany, French, Spanish, etc.
SynthText [23]	No	No	800,000	800,000	-	English
DOST (this paper)	Yes	Yes	32,147	797,919	22,398	Japanese, etc

[a] The number of digits is shown

3 Construction of DOST Dataset

DOST dataset was constructed through the following procedure.

1. Image capture
   Scene images were captured with an omnidirectional camera, Point Grey Ladybug3, consisting of five horizontal and one upward cameras shown in Fig. 1. It was set up on a cart shown in Fig. 2 with a laptop computer and a battery for car. A pair of students walked in a shopping street putting the



☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ ワ
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ #Don'tCare#
☑ 天然石
☑ 天

Fig. 8. Ground truthing software that can transfer text information (label) to neighboring frames.

cart. Images were captured in 6.5 fps in the uncompressed mode. The resolutions of each captured image were $1,200 \times 1,600$ pixels. Lens distortion of the captured images was rectified by a provided software by the vendor of the camera. This process completed in the year of 2012. Table 3 summarizes where, how long and how many images we captured.

2. Ground truthing
   Selected sequences were ground truthed by hand, unlike COCO-Text dataset [29] that used existing scene text detection/recognition methods. The reasons we did not use these methods were that scene texts contained in these images were very difficult for these methods. We developed a ground truthing tool shown in Fig. 8 to make it efficient. Similar to LabelMe Video [30], it had a functionality to transfer text information (text label) in a frame to neighboring frames using homography. However, things in the scene were not on a plane as homography assumes. Hence, following homography computation, more precise positions of words were determined by sliding window based template matching. Table 4 shows distribution of lengths of sequences. Each image is checked at least twice by different persons; one for ground truthing and the other for confirmation. When the ground truthing policy is updated, ground truths are updated by the confirmation opportunity. We spent more than 1,500 man hours for this process.
3. Privacy preservation
   Since the captured images preserved the real scene in shopping streets, we cannot avoid capturing passengers. To avoid privacy violation, we blurred face regions of passengers. At first, we used Amazon Mechanical Turk service. Later, however, we decided to ask this task also to our students so as to ensure the quality with less managing efforts.

Table 3. Place, time length (in hour), the number of images of capture.

Place	Length [h]	#Image
Sakai-Higashi	0.73	101,874
Namba	3.71	521,988
Shinsaibashi	0.25	35,100
Abiko	0.50	70,614
Tennoji	0.38	53,754
Total	5.57	783,150

Table 4. Distribution of lengths of image sequences.

Length of sequence	#sequence
5001	2
3181	1
2000 – 2009	4
1951	1
1500 – 1501	2
101 – 582	6
–100	9
Total	27

4 Ground Truthing Policy

The ground truthing policy of DOST dataset is almost shared with the 2013 and 2015 editions of ICDAR Robust Reading Competition Challenge 3 datasets [5,24]. Since DOST dataset contained not only Latin but also Japanese text, in addition to the ground truthing policy for Latin scripts, we determined one for Japanese text. The ground truthing policy of DOST dataset is summarized below.

1. Basic unit
   A bounding box is created for each basic unit such as a word. In Latin text, word regions segmented by a space is a basic unit. On the other hand, a Japanese sentence is written without some space between words or grammatical units. Hence, as a basic unit of a Japanese sentence, we use *bunsetsu* which is the smallest unit of words that sounds natural in a spoken sentence. A proper noun is not divided.
   There is an exception. If the quality of text is "low," multiple texts of low quality are covered by a single bounding box (see "Transcription" below).
2. Partial occlusion and out of frame
   Even if the region of a basic unit is partially occluded or partially out of frame, it is regarded as a single basic unit without division.
3. Bounding box
   To cope with perspective distortion, a bounding box of a basic unit is represented by four isolated points.
4. Transcription
   The transcription of a basic unit region consists of visible characters. If a basic unit region is partially occluded or partially out of frame, visible characters are transcribed and invisible character(s) are represented by a space. For example, there is a segmented word region of "Barcelona" but "ce" is occluded. Then, the transcription should be "Bar lona." In Fig. 6, an underscore represents a space.

5. ID
   The same ID is assigned to a sequence of a basic unit as long as it can be traced within the frame. An exception is the case a basic unit once completely disappears because it goes out of the frame; in such a case, even if it appears again, a different ID is assigned to the new one.
6. Quality
   Either "high," "medium" or "low" is assigned to each basic unit based on subjective evaluation. Basic units with "high" and "medium" are regarded as legible. We allowed to enlarge the image to check if they are legible. Basic units with "low" are regarded as "don't care" regions where even if a text detection method detects such basic units, it is not considered as failure in detection.
7. Language
   Either "Latin" or "Japanese" is assigned to each basic unit. A basic unit consisting of only alphabets and digits is labeled as "Latin." A basic unit containing at least one non-alphabet or non-digit character is labeled as "Japanese." This is useful for performing an experiment using only Latin text.

5 Comparison of Datasets

Difficulty of major datasets were compared using two detectors and one end-to-end recognition method. To reduce computational burden, in some datasets, a part of data were randomly sampled and used for the experiment. The datasets compared and how they were processed were described below.

1. ICDAR2003 [4]
   All (258) images in the training set were used in the experiment.
2. ICDAR2013 (Challenge 2)[5]
   All (229) images in the training set were used.
3. ICDAR2015 (Challenge 3) [24]
   Images were sampled once in every 30 frames in 10 out of 24 training videos. As a result, 207 images were selected.
4. ICDAR2015 (Challenge 4) [24]
   All (1,000) images in the training set of "End to End" task (Task 4.4) of ICDAR 2015 Robust Reading Competition Challenge 4 were used.
5. SVT [3]
   All (350) images in both training and test sets were used.
6. YVT [25]
   Images were sampled once in every 30 frames in all (30) videos. As a result, 420 images were selected.
7. COCO-Text [29]
   300 images were randomly sampled from ones containing words annotated as English, legible and machine printed (say, target words). The 300 images contained 2,403 target words and words which do not satisfy the condition of the target words (say, non-target words). The non-target words were treated as "don't care" regions.

8. DOST (this paper)
   Images were sampled once in every 30 frames in all ground truthed sequences. As a result, 1,075 images were selected.
9. DOST Latin (this paper)
   This is to evaluate DOST dataset as a Latin scene text dataset containing only alphabets and digits. In text detection and recognition, the same images as "DOST" presented above were used. In evaluation, words containing characters other than alphabets and digits were treated as "don't care" regions. Thus, even if Japanese texts are detected, it does not affect the result.

Two detection methods were used for evaluation. One was the scene text detection method contained in the OpenCV API version 3.0. It was based on Neumann et al. [31]. The other was Matsuda et al. [32]. We were privately given the source code by courtesy of the authors of the paper. In addition, Google Vsion API[2] was used as a powerful commercial end-to-end recognition system. We could designate the language of texts. Only for "DOST," we designated Japanese. In this mode, English texts are also able to be detected and recognized while accuracies are expected to be lower. For other datasets including "DOST Latin," we designated English.

In performance evaluation, regardless of datasets, we shared the same evaluation criteria. For both text detection and end-to-end word recognition tasks, we followed the evaluation criteria used in the challenge of "incidental scene text" (Challenge 4) of ICDAR 2015 Robust Reading Competition. That is, for the scene text detection task, based on a single Intersection-over-Union (IoU) criterion with a threshold of 50 %, a detected bounding box was regarded as correct if it overlapped by more than 50 % with a ground truth bounding box. Recall

Table 5. Detection and Recognition results on selected datasets. Evaluation criteria are recall (R), precision (P) and F-measure (F) in percentage.

Dataset	Text detection						End-to-end		
	OpenCV API			Matsuda [32]			Google Vision API		
	R	P	F	R	P	F	R	P	F
ICDAR2003 [4]	17.6	20.0	21.1	35.0	74.2	47.5	77.3	86.1	81.8
ICDAR2013 (Challenge 2) [5]	11.4	4.2	6.1	4.8	5.2	4.8	70.9	71.8	71.3
ICDAR2015 (Challenge 3) [24]	9.7	7.5	8.5	2.4	10.4	3.9	38.2	52.0	44.1
ICDAR2015 (Challenge 4) [24]	11.4	15.1	13.0	3.8	18.1	6.3	40.5	61.6	48.5
SVT [3]	26.3	14.9	19.0	27.6	30.6	29.1	31.5	19.6	24.2
YVT [25]	36.4	23.4	28.5	1.1	5.3	1.9	33.1	43.8	37.7
COCO-Text [29]	9.3	16.5	11.9	0.8	11.3	1.5	11.9	30.5	17.1
DOST (this paper)	1.4	9.4	2.4	1.6	14.1	2.8	1.7	6.5	2.7
DOST latin (this paper)	0.8	2.2	1.2	1.3	5.2	2.1	6.6	39.6	11.2

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

and precision were simply calculated by the following equations.

$$\text{Recall} = \frac{\text{Number of correctly detected bounding boxes}}{\text{Number of bounding boxes in ground truth}} \tag{1}$$

$$\text{Precision} = \frac{\text{Number of correctly detected bounding boxes}}{\text{Number of detected bounding boxes}} \tag{2}$$

Then, F-measure was calculated as the harmonic mean of precision and recall. For the end-to-end word recognition task, a detected bounding box was regarded as correct if it satisfies the condition of the scene text detection task as well as the estimated transcription was completely correct. Recall, precision and F-measure were calculated in the same way as the detection task.

Results are summarized in Table 5. As can be seen, results of "DOST" and "DOST Latin" were far worse than others. This indicates that DOST dataset reflecting the real environment is more challenging than the major benchmark datasets.

6 Conclusion

Although many scene text datasets publicly available already exist, none of them are intentionally constructed to reflect the real environment. Hence, even though scene text detection/recognition methods achieve high accuracies on these existing major benchmark datasets, it was not possible to evaluate how they are good for practical use. To address the problem, we presented a new scene text dataset named Downtown Osaka Scene Text Dataset (in short, DOST dataset). Unlike most of existing datasets consisting of scene images intentionally captured, DOST dataset consists of uncontrolled scene images; use of an omnidirectional camera enabled us to capture videos (sequential images) of whole scenes surrounding the camera. Since the dataset preserved the real scenes containing texts as they were, in other words, they are *scene texts in the wild*. Through the evaluation conducted in the paper to know the difficulty and quality in comparison with existing datasets, we demonstrated that DOST dataset is more challenging than the major benchmark datasets.

Acknowledgments. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by JST CREST and JSPS KAKENHI #25240028.

References

1. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of CVPR, pp. 4168–4176 (2016)
2. Mishra, A., Alahari, K., Jawahar, C.V.: Scene text recognition using higher order language priors. In: Proceedings of BMVC (2012)
3. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: Proceedings of ICCV, pp. 1457–1464 (2011)

4. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H., Miyao, H., Zhu, J., Ou, W., Wolf, C., Jolion, J.M., Todoran, L., Worring, M., Lin, X.: ICDAR 2003 robust reading competitions: Entries, results and future directions. IJDAR **7**(2–3), 105–122 (2005)
5. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Gomez i Bigorda, L., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., de las Heras, L.P.: ICDAR 2013 robust reading competition. In: Proceedings of ICDAR, pp. 1115–1124 (2013)
6. Wang, T., Wu, D.J., Coates, A., Ng, A.Y.: End-to-end text recognition with convolutional neural networks. In: Proceedings of ICPR, pp. 3304–3308 (2012)
7. Novikova, T., Barinova, O., Kohli, P., Lempitsky, V.: Large-lexicon attribute-consistent text recognition in natural images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 752–765. Springer, Heidelberg (2012)
8. Goel, V., Mishra, A., Alahari, K., Jawahar, C.V.: Whole is greater than sum of parts: recognizing scene text words. In: Proceedings of ICDAR, pp. 398–402 (2013)
9. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: Photoocr: reading text in uncontrolled conditions. In: Proceedings of ICCV, pp. 785–792 (2013)
10. Alsharif, O., Pineau, J.: End-to-end text recognition with hybrid HMM maxout models. In: International Conference on Learning Representations (ICLR) (2014)
11. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Word spotting and recognition with embedded attributes. IEEE TPAMI **36**(12), 2552–2566 (2014)
12. Yao, C., Bai, X., Shi, B., Liu, W.: Strokelets: a learned multi-scale representation for scene text recognition. In: Proceedings of CVPR (2014)
13. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 512–528. Springer, Heidelberg (2014)
14. Su, B., Lu, S.: Accurate scene text recognition based on recurrent neural network. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9003, pp. 35–48. Springer, Heidelberg (2015)
15. Rodriguez, J.A., Gordo, A., Perronnin, F.: Label embedding: a frugal baseline for text recognition. IJCV **113**(3), 193–207 (2015)
16. Gordo, A.: Supervised mid-level features for word image representation. In: Proceedings of CVPR (2015)
17. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. IJCV **116**(1), 1–20 (2016)
18. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Deep structured output learning for unconstrained text recognition. In: Proceedings of ICLR (2015)
19. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. CoRR abs/1507.05717 (2015)
20. Poznanski, A., Wolf, L.: CNN-N-gram for handwritingword recognition. In: Proceedings of CVPR (2016)
21. Liang, J., Doermann, D., Li, H.: Camera-based analysis of text and documents: a survey. IJDAR **7**(2), 83–104 (2005)
22. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. In: Proceedings of NIPS Deep Learning Workshop (2014)
23. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of CVPR (2016)

24. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., Shafait, F., Uchida, S., Valveny, E.: ICDAR 2015 robust reading competition. In: Proceedings of ICDAR, pp. 1156–1160 (2015)
25. Nguyen, P.X., Wang, K., Belongie, S.: Video text detection and recognition: dataset and benchmark. In: Proceedings of WACV (2014)
26. Nagy, R., Dicker, A., Meyer-Wegener, K.: NEOCR: a configurable dataset for natural image text recognition. In: Iwamura, M., Shafait, F. (eds.) CBDAR 2011. LNCS, vol. 7139, pp. 150–163. Springer, Heidelberg (2012)
27. Jung, J., Lee, S., Cho, M.S., Kim, J.H.: Touch TT: scene text extractor using touchscreen interface. ETRI J. **33**(1), 78–88 (2011)
28. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
29. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: COCO-Text: dataset and benchmark for text detection and recognition in natural images. CoRR abs/1207.0016 (2016)
30. Yuen, J., Russell, B., Liu, C., Torralba, A.: LabelMe video: building a video database with human annotations. In: Proceedings of ICCV, pp. 1451–1458 (2009)
31. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: Proceedings of CVPR, pp. 3538–3545 (2012)
32. Matsuda, Y., Omachi, S., Aso, H.: String detection from scene images by binarization and edge detection. Trans. IEICE **J93**(3), 336–344 (2010). In Japanese