RoadText-1K: Text Detection & Recognition Dataset for Driving Videos

Sangeeth Reddy[1], Minesh Mathew[1], Lluis Gomez[2], Marçal Rusinol[2], Dimosthenis Karatzas[2] and C.V. Jawahar[1]

***Abstract*— Perceiving text is crucial to understand semantics of outdoor scenes and hence is a critical requirement to build intelligent systems for driver assistance and self-driving. Most of the existing datasets for text detection and recognition comprise still images and are mostly compiled keeping text in mind. This paper introduces a new "RoadText-1K" dataset for text in driving videos. The dataset is 20 times larger than the existing largest dataset for text in videos. Our dataset comprises 1000 video clips of driving without any bias towards text and with annotations for text bounding boxes and transcriptions in every frame. State of the art methods for text detection, recognition and tracking are evaluated on the new dataset and the results signify the challenges in unconstrained driving videos compared to existing datasets. This suggests that *RoadText-1K* is suited for research and development of reading systems, robust enough to be incorporated into more complex downstream tasks like driver assistance and self-driving. The dataset can be found at `http://cvit.iiit.ac.in/research/projects/cvit-projects/roadtext-1k`**

I. Introduction

Recently, advanced driver assistance and self driving systems have become an active research area. Current driver assistance systems and self-driving approaches mostly disregard textual information on road although text is a medium for conveying crucial information to human drivers. Autonomous navigation systems rely on information from maps, sensory and visual feed [1], [2], [3] for route planning and safe navigation on road. Most of the existing driving datasets [4], [5], [6] have pixel level annotations for objects and other semantic aspects, but do not have text annotated. However, in a real driving environment, it is very common to have unanticipated situations on road leading to interim diversions and detours from regular regime. Generally in such cases text warning boards are used as medium of communication to the driver. Such situations make it necessary for the system to understand the text and act accordingly. A few of such instances where cognizance of text is of paramount importance is depicted in the Fig. 1. A new dataset for this purpose is introduced motivating to have systems that are capable of detecting and recognizing text precisely on road, equipping the relevant systems with textual information in real time.

Scene Text detection and recognition has drawn a lot of interest in the computer vision community for its applications in varied domains, ranging from aiding visually impaired individuals to image search and retrieval. With the advent of deep learning and abundance in digital data there has been considerable progress in scene text detection and recognition. Though the application of scene text in images is well explored for retrieval [7], [8], [9], fine grained classification of products [10], [11] and businesses [7], [11], text translation [12] and tools for the visually impaired [13], it has not been completely incorporated into driver assistance systems or self-driving, except for sign board detection [14], [15].

Fig. 1: Examples showing the importance of text for self-driving and advanced driver assistance systems. Navigation systems in the case of scene on the left, without cognizance of text would be taking the road while it is not supposed to. Similarly the scene at the top-right shows no right turn, but the restriction is only for the timings mentioned. The image on the bottom-right has one way sign to the right but has conflicting message through text.

Most of the recent advances in text detection and recognition in natural scenes [16], [17], [18], [19], [20] deal with only static frames. It is often very challenging to detect and recognize text in video frames due to various factors like blur, out-of-focus and other artifacts/distortions as depicted in Fig. 2. There has been recent interest [21], [22] in community for extending text detection and recognition to videos and there are a handful of datasets [23], [24], [25] that support research in this domain. Existing video text datasets and ICDAR 2013, Robust Reading Challenge [26] are text centric and curated specifically for this purpose. In the case for autonomous navigation, text present in driving environment is widely spread on scene and camera need not necessarily be centered on text. Adding to it, camera movement incorporates artifacts like motion blur. Hence the techniques built on existing datasets which have focused and centered text, are quite not a match for real world applications such as driver assistance and self-driving systems. Our work intends to contribute a dataset RoadText-1K to the community for developing and testing systems that can fare in realistic

[1]The authors are with Center for Visual Information Technology (CVIT), IIIT Hyderabad, India. sangeeth.battu@research.iiit.ac.in

[2]The authors are with Computer Vision Center (CVC), UAB, Spain. lgomez@cvc.uab.es

(a) Motion blur (b) Low contrast (c) Out-of-focus (d) Glare

Fig. 2: Frames from RoadText-1K illustrating various challenges/artefacts often encountered in driving videos.

settings.

Contributions of this work are,

- We create a large scale, diverse and unconstrained dataset of driving videos with dense annotations of text location and transcription. The proposed dataset is 20 times the existing largest dataset [26].
- License plates in the videos are separately tagged to distinguish them from other text instances. This would make the data useful for the problem of license plate detection and recognition.
- We evaluate current state of the art techniques for scene text detection, recognition, tracking and provide a thorough analysis of performance on this dataset.

Rest of the paper is structured as follows, Section II discusses the related work concerning text detection and recognition in images and videos. Section III details how new dataset is compiled, annotated and presents statistics, comparative analysis. In Section IV we present results of state-of-the-art text detection, recognition and tracking methods on the new dataset. Finally, Section V presents the conclusion and thoughts for future work.

II. Related Work

In this section we discuss various related works concerning detection and recognition of text in images and videos, followed by a discussion of works which make use of text in scenes for other applications and downstream tasks.

A. Text detection and recognition in scene images

Text detection and recognition has been an interesting field of research for a long time. With the advent of deep learning and success of Convolutional Neural Networks (CNN), most of the recent text detection and recognition methods have been exploiting their share of utility from these learning models. Strong performance of CNN for object detection tasks has motivated the community to use them for text detection. Text being treated as an object in the image, various methods [27][28][29] based on object detection architectures as backbone have been proposed and have improved the performance considerably. On the other hand, most recent methods for text recognition i.e, the task of transcribing text in a localized region in an image, treat the problem as a sequence to sequence translation problem and primarily rely on a Convolutional Recurrent Neural Network (CRNN) style architecture [30] or an encoder-decoder approach [31].

B. Text detection and recognition in videos

In general, video frames particularly of an outdoor video involving motion, is subject to various artefacts like motion blur and defocus Methods designed for still images, may fail to obtain reliable detection and recognition results when applied to frames of a video. On the other hand a text instance appears in multiple frames in a video and this temporal redundancy can be of help in improving the recognition. Exploiting the redundancy and correlation of textual features across temporal domain is expected to improve the detection and recognition results compared to single frame level methods. Various video text detection methods explore this strategy [32], [33], [34], [35], [36] by techniques such as tracking, multi-frame integration and spatio-temporal analysis. Yin et al. [37] summarizes text detection, tracking and recognition methods in video and their challenges.

Once the text regions are tracked using various tracking methods, one of these two techniques are typically employed for better recognition:(1) *(selection technique)* by selecting best text instances from tracked text regions as proposed in [38] [34], and (2) *(fusion technique)* by combining consecutive recognition results. Rong et al. [39] fuse multiple recognition results of same text region and final result is considered by either majority voting or fusing based on confidence or other metrics. Nevertheless these approaches are computationally expensive since the detection and recognition models essentially run on every frame.

Wang et al. in [40] propose a method for end-to-end text recognition utilizing correspondences across multiple frames. They use spatio-temporal redundancy for text detection and edit distance for recognized text to fuse results from multiple frames. Recent work from Cheng et al. [41], propose a scene text spotting framework in video with a spatiao-temporal detector and discriminative tracker to recognize text once per track by picking the best instance using a quality scoring mechanism. In section V we evaluate a similar strategy for

Fig. 3: Example frames from clips in RoadText-1K with text location and transcription annotations overlaid. Boxes in green correspond to English text, blue represent non English and red represent illegible text.

tracking on the proposed RoadText-1K dataset.

C. Scene text understanding for navigation

Autonomous navigation systems highly rely upon visual and sensory inputs. A camera feed is used commonly for semantic segmentation or classification of scene into drivable area, sidewalk, vegetation, sky etc. However the text present in the scene which conveys high level semantic information, is often ignored.

Case et al. [42] propose to use text present in the scene to generate automatic semantic labels during robotic mapping of an indoor environment. Work by Wu et al. [14] deals with text detection on road signs with application to driver assistance systems. In this work the sign boards are first detected in each frame of the video and then features like color, edges and texture are used for detection of text on the sign boards. Gonzalez, et al. [43] attempts to extract the text present in road panels of street images as an application to Intelligent Transportation Systems (ITS). A character and word recognizer were used for recognition and to improve the recognition, a Web Map Service is used to restrict the dictionary size to a limited geographical area. Shi et al. [44] propose to detect and recognize text in traffic signs by using Maximally Stable Extremal Regions (MSER), hue, saturation, value color thresholding with constraints upon temporal and structural information for text region detection. Individual text characters are detected as MSER and grouped into lines followed by an Optical Character Recognition (OCR) module. Zhou et al. [28] detects text-based traffic signs using a convolutional network as region proposal network and another neural network for final classification of text regions from the proposals.

Our work goes beyond traffic-sign text detection, and introduces a new large-scale, densely annotated dataset and benchmark for scene text detection and recognition in an unconstrained, realistic driving setup.

III. RoadText-1K

We start by explaining how videos for the dataset were selected from an existing driving videos dataset followed by annotation process. Finally, we provide statistics and analysis of the dataset in comparison with existing datasets for text in videos

A. Videos

The 10 second long video clips in our dataset are sampled from BDD100K [45], which contain 100K driving videos. Each video in BDD100K is about 40 seconds long, 720p and 30 fps collected with diversity across locations in the United States, weather conditions including sunny, overcast, and rainy, as well as different times of day including day and night. Videos in BDD100K are collected with an intent to make robust self driving systems and were not specifically collected for the task of text detection or any aligned task. This makes BDD100K video database an ideal source of raw videos that are diverse and unconstrained without any specific bias towards the problem of understanding text in images/videos.

We ran an off-the-shelf text detector [28] on frames of the videos in BDD100K to shortlist videos which have considerable number of text instances. In the next step, 1000, 10 seconds long video clips are handpicked from this shortlisted collection. We performed this step manually to make sure that the clips selected are diverse in terms of scenes, and have good number of English text. We split the 1000 videos in the dataset randomly as 700 for training and 300 for test.

B. Annotation of text instances

The annotation of text instances in every frame of the video clips involved a two stage process. In the first stage the annotators were asked to bound the text instances by a bounding box and to assign each of them with a text type category from the following three categories: (1) English (2) Non English (3) Illegible. In English category we further classify into (a) English or (b) license plate. Distribution of text instances in RoadText-1K based on the text type is shown in Fig. 5.

Unlike in the case of most scene text datasets we annotate text lines rather than annotating each ‘word’ (split at spaces). This approach makes annotation much faster and also avoids the ambiguity in deciding how to split text into ’words’ in certain cases where numbers or abbreviations are involved. Also when text is annotated at ‘word’ level, it will result in many smaller tokens like a single full stop, a single ‘a’ or a single digit which are typically difficult to detect and recognize. Moreover the recent text recognition methods using Connectionist Temporal Classification (CTC) [46] or an encoder-decoder framework doesn’t require the text line to be split into words or sub-words. Also transcribing a text line at once lets the transcription model access more context, and can benefit from an implicit language model which is learnt as part of the process [47].

(a) English (b) Illegible (c) Non English

Fig. 4: Samples of text instances for each text category

Dense annotation of text in video is a complex and time consuming task, since it involves annotating bounding box for every text instance and at every frame. The major advantage of video is to have multiple occurrences of same text temporally and with considerably small change in spatial locations depending on the motion. Hence we used Scalabel[1] for bounding box annotation as it enables tracking across frames. Once a text instance is marked with a bounding box, a track id and category label, the tool aids in tracking the instance across frames and assigns the same track id and category label to other occurrences of the same text instance. The human annotator then reviews the boxes frame by frame and manually adjusts the boxes wherever tracking did not fit the boxes correctly. There were cases where a text instance undergo transition from illegible/occluded to a legible case or vice-versa. In such cases, initial track is ended from the point where the transition happens and a new track is started, with a new category label. And in cases where a legible text gets partially occluded or goes out of focus for a few frames in between, annotators were directed to continue the track if the occlusion/out-of-focus is not for more than 3 frames. If it takes more than 3 frames, existing track is ended, and a new track with category label as *illegible* is assigned for the time the text instance is occluded/out-of-focus.

In the second stage of annotation, annotators were given videos annotated with the bounding boxes in the first step and were asked to add text transcriptions for tracks which are assigned with *English* and *license plate* category. For this step we used a tool developed internally for this task. The tool loads a video with the bounding box annotations and displays track ids and the category labels assigned in the first step. Once the annotator adds the text transcription to a certain text instance, it is propagated to all occurrences in the same track. Hence the annotator needs to attend to only those tracks which are of *English* or *license plate* category and not assigned with a transcription yet. Examples of frames with both the bounding box and transcription annotations overlaid are shown in Fig. 3.

C. Statistics and Analysis

We summarize the overall statistics of RoadText-1K in Table I along with a comparison of our dataset with existing datasets for text in videos. Top 10 frequent words in our dataset are shown in Fig. 7(a), we also plot the distribution of words as a function of word lengths in Fig. 7(b) . It is observed from the plot that most words present on road are in general of length 7-8 characters and can go till 41 characters. Fig. 7(c) shows number of frames as a function of number of

[1]https://www.scalabel.ai

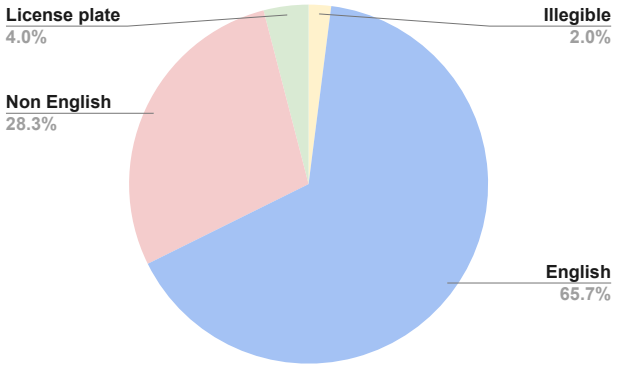


Fig. 5: Category distribution of text instances in RoadText-1K.

text instances appearing in a frame. Although the dataset is collected on roads, at least 50% of the unique text instances are non traffic/road signs.

Existing datasets we compare in Table I are : Text in Videos [48], USTB-VidTEXT [23] and Youtube Video Text (YVT) [25]. Among these, USTB-VidTEXT and YVT mostly contain born digital text (captions and subtitles) in videos sourced from Youtube. Recognition of born digital text is less challenging since they are free from most of the distortions and imaging artifacts and hence even OCRs designed for documents perform quite well on them. Text in Videos dataset is the largest among these and have 50 egocentric videos containing mostly scene text with dense annotations for bounding boxes and transcriptions. In Fig. 8 we show the spatial distribution of text in these datasets compared to RoadText-1K. It can be seen that both RoadText-1K and Text in Videos datasets have text instances spread widely across the frame compared to the other two. And the spread in USTB-VidTEXT is very minimal since it only contains subtitles.

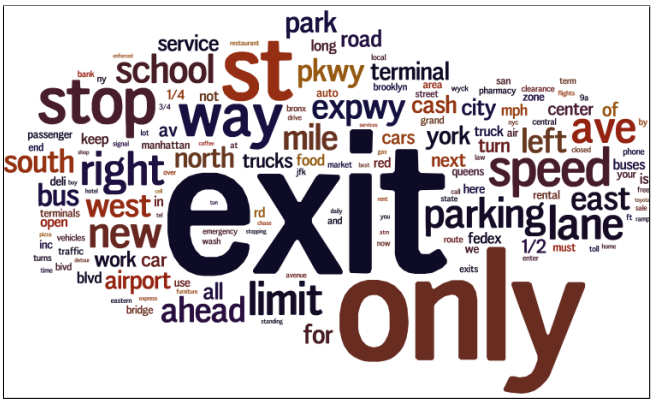
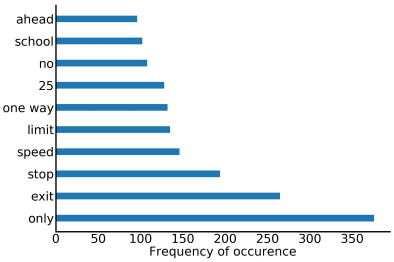


Fig. 6: WordCloud of words in RoadText-1K

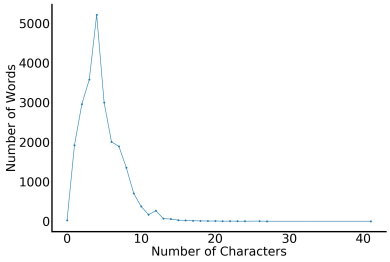
Apart from video based datasets, Uber-Text [49] is large

TABLE I: Comparison of RoadText-1K with existing datasets for text in videos

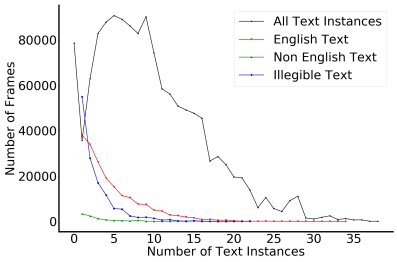
Dataset	Text in Videos [48]	USTB-VidTEXT [23]	YouTube Video Text [25]	RoadText-1K (ours)
Source	Egocentric	Youtube	Youtube	car-mounted
Size (Videos)	51	5	30	**1000**
Length (Seconds)	varying	varying	15	10
Resolution	720 × 480	480 × 320	1280 × 720	1280 × 720
Annotated Frames	27,824	27,670	13,500	**300,000**
Total Text Instances	143,588	41,932	16,620	**1,280,613**
Text type	Scene Text	Digital (captions)	Scene Text and Digital	Scene Text
Unique Words	3,563	306	224	8,263
Avg. text frequency per frame	5.1	1.5	1.23	4.2
Avg. Text Track length	46	161	72	48



(a) Top 10 most occurring words.

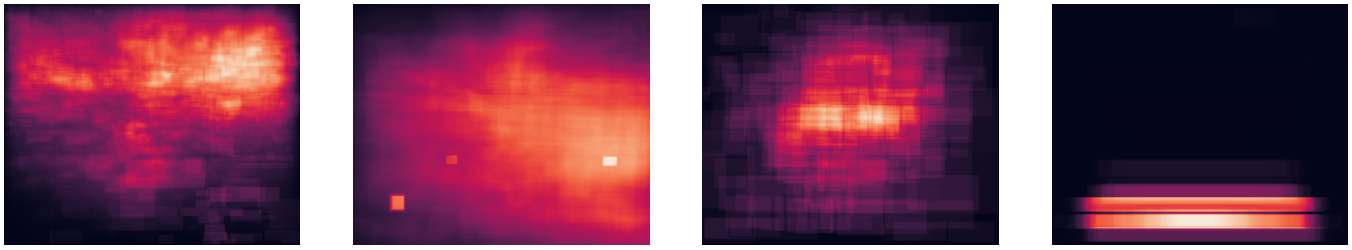


(b) Number of words with a particular length.



(c) Number of frames with a particular number of text instances.

Fig. 7: Statistics of text instances in RoadText-1K.



(a) RoadText-1K (b) Text in Videos (c) Youtube Video Text (d) USTB-VidText

Fig. 8: Heatmaps depicting the spatial distribution of text instances for existing and proposed datasets.

scale image dataset closest to the proposed RoadText-1K. This dataset is collected from images by the Bing Maps Streetside program largely aligning with our unconstrained driving videos from BDD100K. UberText contains 117969 street-level images with 571534 annotated text regions. The dataset also categorize text instances into 9 categories indicating whether the text is business name, street name etc. UberText aligns with the proposed dataset in terms of non text centered/focused frames, while the differences are in the bounding box shape. We provide rectangular bounding boxes while UberText provides free-hand polygon annotations for text instances.

IV. Experiments

A. Text Detection and Recognition in Images

Various existing techniques for text detection and recognition are evaluated on the RoadText-1K, such as CTPN [27], EAST [28] and FOTS [29] for detection and CRNN [50] and ASTER [31]. CRNN is quite popular for introducing the sequence to sequence model for text recognition and had significant performance improvement [50] when proposed, while ASTER is currently state of the art in many popular datasets [31]. Since these techniques are developed for image, we generate still images from videos and the frame level annotations serve as its ground truth.

The Connectionist Text Proposal Network (CTPN) model uses VGG16 backbone for feature extraction, followed by a Bi-directional LSTM for output. Efficient and Accurate Scene Text Detector (EAST) is another state of the art text detector that uses fully convolutional layers for text prediction followed by Non-Maximum Suppression (NMS). Fast Oriented Text Spotting (FOTS) an end-to-end model, shares convolutional layer's features for detection and recognition. The pretrained models of discussed text detection methods CTPN, EAST and FOTS have been trained on word images. While the current annotations are at line level making it not really a fair direct comparison. Hence we finetune the three models on the train set and evaluate on test set of RoadText-1K. Table II below, shows the detection results.

As mentioned earlier we use the existing CRNN, ASTER methods to evaluate recognition performance on frames. CRNN is one of the most commonly used method for text

TABLE II: Frame level text detection results of existing models on the RoadText-1K.

Method	Precision	Recall	F-score
CTPN	0.44	0.41	0.42
EAST	0.42	0.30	0.35
FOTS	0.45	0.36	0.40

recognition which uses a CNN to extract features followed by a Bi-directional LSTM layer for modelling the sequence and finally CTC loss for training. ASTER first uses a spatial transformer network [51] to rectify the images followed by an encoder-decoder style recognition network to transcribe the rectified image. These methods are evaluated on the new dataset and the results are shown in Table III

TABLE III: Frame level text recognition accuracy of existing models on RoadText-1K, given ground truth text line crops and tracks.

Method	Pretrained			Fine tuned		
	All	AN	MV	All	AN	MV
CRNN	29.0	44.6	60.1	36.3	50.9	65.2
ASTER	44.6	61.9	67.2	48.1	63.0	68.3

In the Table III, category "ALL" refers to text instances which are of English legible category including special characters. While "AN" refers to alphanumeric which includes only alphanumeric characters, this is done to provide a fair comparison as pretrained models are in general trained only on alphanumeric data. "MV" refers to Majority Voting, where we consider the text transcription that has maximum occurrence across frames of the same text instance for evaluation.

From Tables II and III we observe that frame level text detection and recognition results on RoadText-1K are not on par with the results these methods report on existing scene text datasets. For example CTPN which performs the best on RoadText-1K , reports an F-score of 0.88 on icdar-2013 [26] benchmark compared to 0.42 on our dataset. Similarly, although the fine-tuned ASTER model yields $60\%+$ accuracy on our dataset, the same model reports $> 90+$ on most benchmark datasets[26], [52] for scene text recognition. The results suggest limitation of the current text detection and recognition methods on unconstrained, realistic driving videos, which are not text centric.

B. Text Detection and Recognition in Videos

We also evaluate video level methods which provide end to end results. The most common metrics used to evaluate are the MOT metrics like MOTP (Multiple Object Tracking Precision), MOTA (Multiple Object Tracking Accuracy). These metrics are widely used in object tracking methods and the same has been used as a metric of evaluation in the ICDAR Robust Reading Challenge (2015) [26] for End to End task in Text in Videos. We evaluate methods built for text tracking in video, object tracking and compare their performance.

The work [40] based on multi frame tracking provides a method to track text instances temporally based on attributes of the text objects in multiple frames. This tracking algorithm considers various factors like IOU, offset of matched frames, edit distance of text to link two text objects across frames to same track. This method has been evaluated with various combinations of text detection and recognition algorithms. The tracking results are presented in the Table IV.

TABLE IV: Text Tracking performance on RoadText-1K using a method proposed in [40]

Method	MOTP		MOTA		ATA %	
	CRNN	ASTER	CRNN	ASTER	CRNN	ASTER
CTPN	17.06	7.4	-29.79	-11	0.56	0.57
EAST	11.48	11.5	-111	-111	0.36	0.31
FOTS	10.75	11	-206	-206	0.13	0.10

In the recent past object tracking community has been using new evaluation metrics for tracking presented in the CVPR19: Tracking and Detection Challenge [53]. These new metrics include *IDF1* (ID F1-score) ratio of correctly identified detections over the average number of ground-truth and computed detections, MT (Mostly Tracked) Number of objects tracked for at least 80 percent of lifespan, ML (Mostly Lost) Number of objects tracked less than 20 percent of lifespan, *Ids* Number of ID switches across tracks, FM Total number of times a trajectory is fragmented along with MOTA Multiple Object Tracking Accuracy, FP False Positives, FN False Negatives. We also evaluate the metrics provided in the challenge using one of the popular object tracking methods SORT [54]. We use detections from various methods and compare the results in Table V.

TABLE V: Evaluation of Text Tracking by SORT [54] using new MOT evaluation metrics proposed in CVPR 2019 Tracking and Detection challenge [53]

Metric	CTPN	EAST	FOTS
MOTA	-65.3	-122.7	-195.0
IDF1 (%)	9.7	12.2	12.15
MT	1.2	1.23	1.82
ML	39.03	23.3	27.8
FP	470	470.7	492.6
FN	628	601.6	654.6
IDs	28.3	22.12	22.89
FM	23.7	19.26	19.36

V. CONCLUSION

We motivate the problem of understanding text in driving videos and introduce a new large-scale dataset for the same. We find that existing datasets for text spotting in images and videos are usually curated with text in mind and hence do not fare well on a realistic setting, like videos captured from a moving car. We believe that RoadText-1K will encourage research both on improving text detection and recognition in videos as well as enabling and equipping complex tasks like self-driving and driver assistance which can hugely benefit from reasoning about text on the road.

REFERENCES

[1] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *CoRR*, 2016.

[2] Z. Chen and X. Huang, "End-to-end learning for lane keeping of self-driving cars," in *IV Symposium*, 2017.

[3] B. Paden, M. Cáp, S. Z. Yong, D. S. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *TIV*, 2016.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *IJRS*, 2013.

[5] G. Varma, A. Subramanian, A. M. Namboodiri, M. Chandraker, and C. V. Jawahar, "IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *WACV*, 2019.

[6] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *CVPR Workshops*, 2018.

[7] S. Karaoglu, R. Tao, T. Gevers, and A. W. M. Smeulders, "Words matter: Scene text for image classification and retrieval," *TMM*, 2017.

[8] A. Mishra, K. Alahari, and C. V. Jawahar, "Image retrieval using textual cues," in *ICCV*, 2013.

[9] L. Gómez, A. Mafla, M. Rusiñol, and D. Karatzas, "Single shot scene text retrieval," in *ECCV*, 2018.

[10] S. Karaoglu, R. Tao, J. van Gemert, and T. Gevers, "Con-text: Text detection for fine-grained object classification," in *TIP*, 2017.

[11] X. Bai, M. Yang, P. Lyu, Y. Xu, and J. Luo, "Integrating scene text and visual appearance for fine-grained image classification," *Access*, 2018.

[12] X. Shi and Y. Xu, "A wearable translation robot," in *ICRA*, 2005.

[13] X. Rong, B. Li, J. P. Muñoz, J. Xiao, A. Arditi, and Y. Tian, "Guided text spotting for assistive blind navigation in unfamiliar indoor environments," in *ISVC*, 2016.

[14] W. Wu, X. Chen, and J. Yang, "Detection of text on road signs from video," *ITS*, 2005.

[15] Y. Yuan, Z. Xiong, and Q. Wang, "An incremental framework for video-based traffic sign detection, tracking, and recognition," *ITS*, 2017.

[16] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *ACCV*, 2011.

[17] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *ICCV*, 2011.

[18] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *TIP*, 2014.

[19] X. Bai, C. Yao, and W. Liu, "Strokelets: A learned multi-scale mid-level representation for scene text recognition," *TIP*, 2016.

[20] L. G. i Bigorda and D. Karatzas, "Textproposals: A text-specific selective search algorithm for word spotting in the wild," *Pattern Recognition*, 2017.

[21] C. Yang, X. Yin, W. Pei, S. Tian, Z. Zuo, C. Zhu, and J. Yan, "Tracking based multi-orientation scene text detection: A unified framework with dynamic programming," *TIP*, 2017.

[22] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "A new technique for multi-oriented scene text line detection and tracking in video," *TMM*, 2015.

[23] S. Tian, X. Yin, Y. Su, and H. W. Hao, "A unified framework for tracking based text detection and recognition from web videos," *TPAMI*, 2018.

[24] R. Minetto, N. Thome, M. Cord, N. J. Leite, and J. Stolfi, "Snooper-track: Text detection and tracking for outdoor videos," in *ICIP*, 2011.

[25] P. X. Nguyen, K. Wang, and S. J. Belongie, "Video text detection and recognition: Dataset and benchmark," in *WACV*, 2014.

[26] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. Almazán, and L. de las Heras, "ICDAR 2013 robust reading competition," in *ICDAR*, 2013.

[27] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *ECCV*, 2016.

[28] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: an efficient and accurate scene text detector," in *CVPR*, 2017.

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

[30] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *TPAMI*, 2017.

[31] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," *TPAMI*, 2018.

[32] L. G. i Bigorda and D. Karatzas, "Mser-based real-time text detection and tracking," in *ICPR*, 2014.

[33] P. Shivakumara, M. Lubani, K. Wong, and T. Lu, "Optical flow based dynamic curved video text detection," in *ICIP*, 2014.

[34] M. Tanaka and H. Goto, "Autonomous text capturing robot using improved DCT feature and text tracking," in *ICDAR*, 2007.

[35] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "A new technique for multi-oriented scene text line detection and tracking in video," *TMM*, 2015.

[36] C. Yang, X. Yin, W. Pei, S. Tian, Z. Zuo, C. Zhu, and J. Yan, "Tracking based multi-orientation scene text detection: A unified framework with dynamic programming," *TIP*, 2017.

[37] X. Yin, Z. Zuo, S. Tian, and C. Liu, "Text detection, tracking and recognition in video: A comprehensive survey," *TIP*, 2016.

[38] H. Shiratori, H. Goto, and H. Kobayashi, "An efficient text capture method for moving robots using DCT feature and text tracking," in *ICPR*, 2006.

[39] X. Rong, C. Yi, X. Yang, and Y. Tian, "Scene text recognition in multiple frames based on text tracking," in *ICME*, 2014.

[40] X. Wang, Y. Jiang, S. Yang, X. Zhu, W. Li, P. Fu, H. Wang, and Z. Luo, "End-to-end scene text recognition in videos based on multi frame tracking," in *ICDAR*, 2017.

[41] Z. Cheng, J. Lu, J. Xie, Y. Niu, S. Pu, and F. Wu, "Efficient video scene text spotting: Unifying detection, tracking, and recognition," *CoRR*, 2019.

[42] C. Case, B. Suresh, A. Coates, and A. Y. Ng, "Autonomous sign reading for semantic mapping," in *ICRA*, 2011.

[43] Á. Gonzalez, L. M. Bergasa, and J. J. Y. Torres, "Text detection and recognition on traffic panels from street-level imagery using visual appearance," *ITS*, 2014.

[44] J. Greenhalgh and M. Mirmehdi, "Recognizing text-based traffic signs," *ITS*, 2015.

[45] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving video database with scalable annotation tooling," *CoRR*, 2018.

[46] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[47] E. Sabir, S. Rawls, and P. Natarajan, "Implicit language model in LSTM for OCR," *CoRR*, vol. abs/1805.09441, 2018. [Online]. Available: http://arxiv.org/abs/1805.09441

[48] X. Zhou, S. Zhou, C. Yao, Z. Cao, and Q. Yin, "ICDAR 2015 text reading in the wild competition," *CoRR*, vol. abs/1506.03184, 2015.

[49] Y. Zhang, L. Gueguen, I. Zharkov, P. Zhang, K. Seifert, and B. Kadlec, "Uber-text: A large-scale dataset for optical character recognition from street-level imagery," in *SUNw: Scene Understanding Workshop - CVPR*, 2017.

[50] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *TPAMI*, 2017.

[51] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *NIPS*, 2015.

[52] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *BMVC*, 2012.

[53] P. Dendorfer, S. H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. D. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "CVPR19 tracking and detection challenge: How crowded can it get?" *CoRR*, vol. abs/1906.04567, 2019.

[54] A. Bewley, Z. Ge, L. Ott, F. T. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *ICIP*, 2016.