Sergi Garcia-Bordils[1,3], George Tom[2], Sangeeth Reddy[2], Minesh Mathew[2], Marçal Rusiñol[3], C.V. Jawahar[2], and Dimosthenis Karatzas[1]

[1] Computer Vision Center (CVC), UAB, Spain
[2] Center for Visual Information Technology (CVIT), IIIT Hyderabad, India
[3] AllRead Machine Learning Technologies

Read while you drive - multilingual text tracking on the road

Motivation

Visual data obtained during driving scenarios usually contain large amounts of text that conveys semantic information necessary to analyse the urban environment and is integral to the traffic control plan. Yet, research on autonomous driving or driver assistance systems typically ignores this information. We present RoadText-3K, a large driving video dataset with fully annotated text to advance research in this direction. RoadText-3K is three times bigger than its predecessor and contains data from varied geographical locations, unconstrained driving conditions and multiple languages and scripts. We offer a comprehensive analysis of tracking by detection and detection by tracking methods exploring the limits of state-of-the-art text detection. Finally, we propose a new end-to-end trainable tracking model that yields state-of-the-art results on this challenging dataset. Our experiments demonstrate the complexity and variability of RoadText-3K and establish a new, realistic benchmark for scene text tracking in the wild.

Overview

Our contributions of this work are the following:

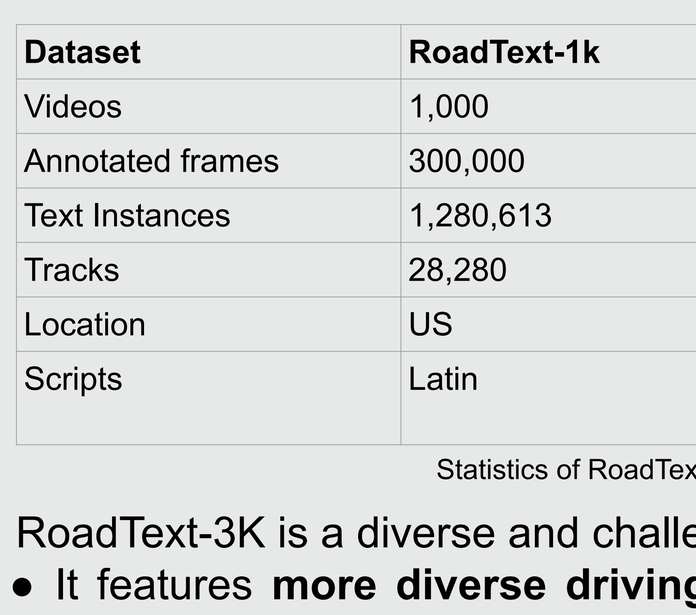
- **An extension of the RoadText-1K**[1] dataset of 2000 more videos, a video driving dataset with fully annotated text. Our dataset features more locations, scripts and driving conditions.
- A **detailed comparative** study between diverse **state-of-art text detection and tracking** models.
- We propose **a new and efficient detection and tracking framework** based on the CenterNet[2] object detector.

Dataset	RoadText-1k	RoadText-3k
Videos	1,000	3,000 (2,000 new videos)
Annotated frames	300,000	927,947
Text Instances	1,280,613	4,039,250
Tracks	28,280	88,427
Location	US	US, Europe and India
Scripts	Latin	Latin, Telugu and Devanagari

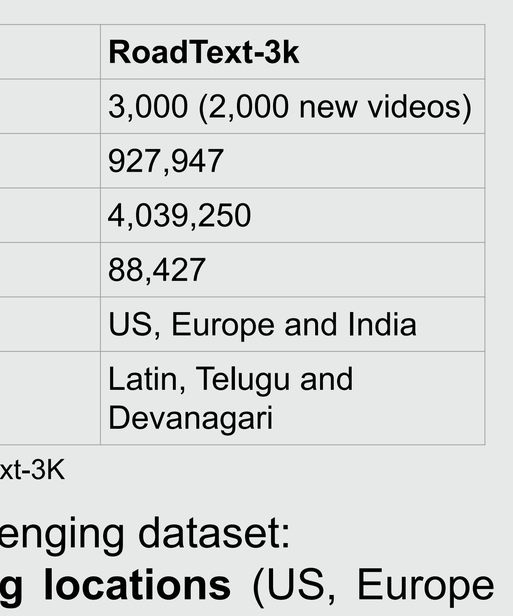
Statistics of RoadText-3K

RoadText-3K is a diverse and challenging dataset:

- It features **more diverse driving locations** (US, Europe and India)
- The videos captured in Europe contain **English, Spanish and Catalan text**.
- The Indian videos not only include English text, but also **different scripts coming from Telugu and Hindi**.



Script and language distribution of RoadText-3K



Sample annotated frames

Baselines and Proposed Framework

In addition to the dataset we evaluate different state-of-art approaches towards text detection and tracking:

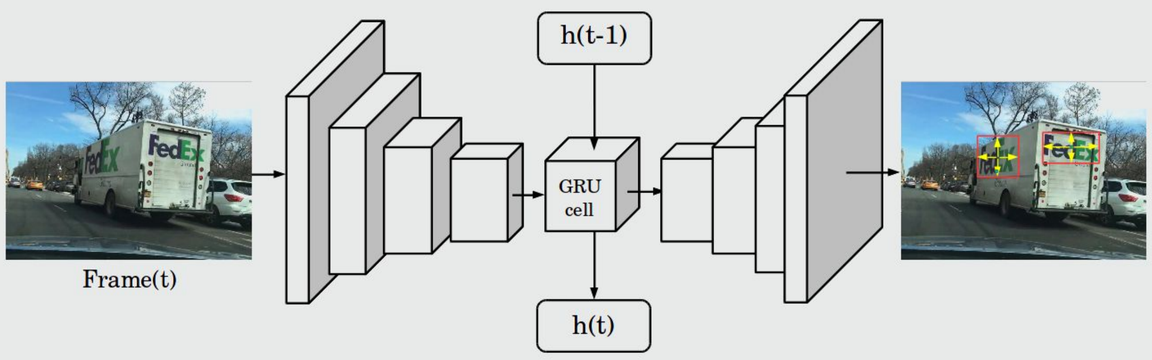
- A **comparison between several state-of-art text detectors** in our challenging dataset.
- A **comparison** between **Tracking by Detection** (TbD) and **Detection by Tracking** (DbT) approaches for text tracking.
- We propose a CenterNet-based text detection and tracking framework tailored for the high-resolution frames of our dataset.

Our CenterNet-based detector offers competitive F-scores at higher FPS than other state-of-art detectors:

Detector	Precision (%)	Recall (%)	F-score (%)	FPS
FOTS	42.77	**50.74**	**46.41**	19
CRAFT	**54.3**	37.21	44.15	5
CenterNet	50.3	39.1	43.9	**44**

Frame-level detection performance on RoadText-3K

The tracking framework uses CenterNet as its detector combined with a convolutional GRU[3] cell to incorporate temporal awareness to our model:



The CenterNet+GRU model **beats the other detectors** used in the metric MOTA, while offering **real time performance**:

Detector	MOTA(%)	IDs	FPS
FOTS	28.47	12883	14
CRAFT	35.40	**6328**	8
CenterNet	33.80	15032	**40**
CenterNet+GRU	**36.00**	8896	31

Tracking performance using TbD

[1] Reddy, Sangeeth, et al. "Roadtext-1k: Text detection & recognition dataset for driving videos." *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.
[2]Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." *arXiv preprint arXiv:1904.07850* (2019).
[3]Cho, Kyunghyun, et al. "On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint arXiv:1409.1259* (2014).

Amb el suport de:
DOCTORATS INDUSTRIALS Alianza de talent!
Generalitat de Catalunya Departament de Recerca i Universitats

Proyecto PID2020-116298GB-I00 financiado por:
MINISTERIO DE CIENCIA E INNOVACIÓN
AGENCIA ESTATAL DE INVESTIGACIÓN