Recognizing Text-based Traffic Guide Panels with Cascaded Localization Network

***Xuejian Rong*[†], *Chucai Yi*[§], *Yingli Tian*[†*]**

[†]*The City College of New York, CUNY, USA;* [§]*HERE North America LLC, USA;* [*]*email: ytian@ccny.cuny.edu.*

Introduction

In this paper, we introduce a new top-down framework for *automatic localization and recognition of text-based traffic guide panels* captured by car-mounted cameras from natural scene images.

Fig. 1: *Samples of traffic guide panels and the text information within.*

Motivation:

- As one of the most important context indicators in driving status, traffic signs (symbol-based or text-based) have attracted considerable attention in the fields of detection and recognition.
- Algorithms for symbol-based traffic signs, e.g. *Stop* or *Exit*, with relatively *smaller size and unique shape*, cannot handle text-based traffic signs/panels with *standard rectangular shape containing extensive text information*.
- Most existing algorithms ignored a large amount of valuable semantic information resided in the text-based traffic signs, which is usually not completely available or up-to-date on car-mounted navigation systems.

Novelties and Contributions: A new top-down framework for automatic localization and recognition of text-based traffic guide panels including:

- A novel Cascaded Localization Network (CLN) joining two customized convolutional nets in the YOLO [23] fashion.
- Popular character-wise text saliency detection is replaced with string-wise text region detection, which avoids numerous bottom-up processing steps.
- A temporal fusion method of text region proposals across consecutive frames.

Approach

Specifically, on a set of continuous image frames captured from the highway environments:

- First the candidate traffic guide panels in each frame are simultaneously extracted using a set of learned convolutional neural network (CNN) features. Input images are evenly divided in a $S \times S$ grid ($S = 7$ for panel detection, $S = 14$ for text detection), and each grid cell is responsible for predicting B bounding boxes and confidence scores.
  Each bounding box is composed of 7 predictions: $\{x, y, w, h, \cos\theta, \sin\theta\}$ and the presence confidence. The (x, y) coordinates and the width/height (w, h) denote respectively the location and size of a bounding box tightly enclosing the guide panel. θ represents the bounding box rotation.
- Then, a fine CNN-based text detector is trained to localize all the text regions within the guide panel candidates.

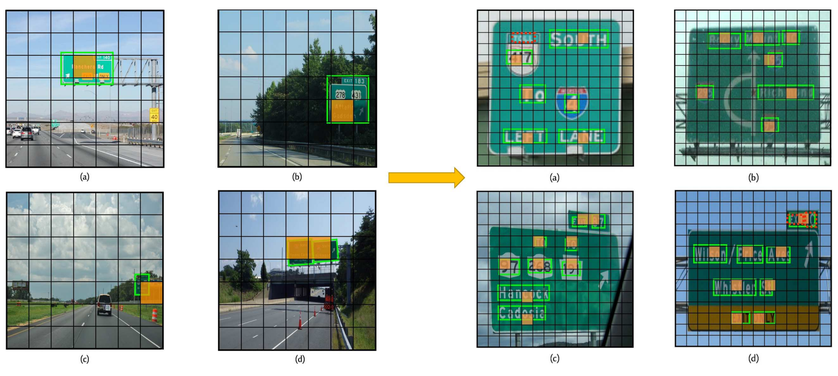

Fig. 2: *Demonstration of localizing guide panel and text region candidates.*

- Third, false positive candidates are eliminated by using temporal information from the continuous frames.
- Finally, these text regions are recognized by a deep recurrent model in a sequence-to-sequence encoder-decoder fashion.

Fig. 3: *The sequence-to-sequence recognition of text region candidates.*

Experiments

Benchmark Dataset: Newly collected dataset which contains a variety of highway guide panels in 3841 high-resolution individual images. All the images are collected from *AAroads* website (`http://www.aaroads.com`).

Comparison to existing methods:

- *Stroke Width Transform*, Epshtein et al. [11]: a well-known method that leverages the consistency of characters' stroke width to detect arbitrary fonts.
- *MSER Text Detection*, Gomez et al. [26]: uses maximally stable extremal regions (MSERs), a popular tool in text detection.
- *Deep Text Spotting*, Jaderberg et al. [8]: a state-of-the-art method that uses multiple stages of convolutional neural networks to predict text saliency score at each pixel, and cluster to form the region predictions afterward.

Table 1: *Text localization results and average processing times on benchmark dataset. Precision P and Recall R at the maximum f-measure F, and the localization time t_l (in seconds).*

Method	P	R	F	t_l
Proposed	**0.73**	0.64	**0.68**	0.16
Jaderberg et al. [8]	0.59	**0.71**	0.64	4.53
Gomez et al. [26]	0.46	0.53	0.49	1.32
Epshtein et al. [11]	0.35	0.48	0.38	2.51

Fig. 4: *Comparison of the Top-5 text region localization proposals from the proposed approach and the best competing baseline method [8].*

Conclusions

- A new top-down CNN-based cascaded framework for automatic detection and recognition of text-based traffic guide panels in the wild.
- Performed in an efficient coarse-to-fine manner, and effectively reduced the redundant computation in continuous frames.

Acknowledgements: This work was supported in part by NSF grants EFRI-1137172, IIP-1343402, and FHWA grant DTFH61-12-H-00002.