Spatially Prioritized and Persistent Text Detection and Decoding

Hsueh-Cheng Wang, Yafim Landa, Maurice Fallon, and Seth Teller

Robotics, Vision, and Sensor Networks Group (RVSN)
Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology (MIT)

Abstract

We show how to exploit temporal and spatial coherence to achieve efficient and effective text detection and decoding for a sensor suite moving through an environment in which text occurs at a variety of locations, scales and orientations with respect to the observer. Our method uses simultaneous localization and mapping (SLAM) to extract planar “tiles” representing scene surfaces. It then fuses multiple observations of each tile, captured from different observer poses, using homography transformations. Text is detected using the Discrete Cosine Transform (DCT) and Maximally Stable Extremal Regions (MSER); MSER enables multiple observations of blurry text regions in a component tree. The observations from SLAM and MSER are then decoded by an Optical Character Recognition (OCR) engine. The decoded characters are then clustered into character blocks, and an MLE word configuration is obtained.

The paper’s contributions include: 1) Spatiotemporal fusion of tile observations via SLAM, prior to inspection, thereby improving the quality of the input data; 2) Scheduling computationally intensive inspection according to a spatial prior on text occurrence, thereby improving efficiency over baseline; and 3) combination of multiple noisy text observations into a single higher-confidence estimate of environmental text.

Keywords: SLAM, Text Detection, Video OCR, Multiple Frame Integration, MSER, DCT, Lexicon, Language Model