Automation in Construction 49 (2015) 27–39

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Segmentation and recognition of roadway assets from car-mounted camera video streams using a scalable non-parametric image parsing method

Vahid Balali [a,*], Mani Golparvar-Fard [a,b,1]

[a] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States
[b] Department of Civil and Environmental Engineering, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, United States

ARTICLE INFO

Article history:
Received 13 March 2014
Received in revised form 6 September 2014
Accepted 30 September 2014
Available online 21 October 2014

Keywords:
Segmentation
Recognition
Parsing
High-quantity low-cost highway assets

ABSTRACT

This paper presents a non-parametric image parsing method for segmentation and recognition of roadway assets such as traffic signs, traffic lights, pavement markings, and guardrails from 2D car-mounted video streams. The method can be easily scaled to thousands of video frames captured during data collection and does not need training. Instead, it retrieves a set of most relevant video frames (e.g. highway vs. secondary road) which serve as candidates for superpixel-level annotation. It then obtains superpixels from the video frames and using the retrieval set encodes their visual characteristics using a histogram of different shape, appearance, and color descriptors. Neighborhood contexts are incorporated by using Markov Random Field (MRF) optimization and two types of semantic (e.g. guardrail) and geometric (e.g. horizontal) labels are simultaneously assigned to the superpixels. We introduce a new dataset from I-57 together with its ground truth and present experimental results on both I-57 and SmartRoad datasets. Experimental results with an average accuracy of 88.24% for recognition and 82.02% for segmentation show that our local visual features provide acceptable performance, while the method overall does not require any significant supervised training. This scalable method has potential to reduce the time and effort required for developing road inventories, especially for those such as guardrails and traffic lights that are not typically considered in 2D asset recognition methods.

1. Introduction

High quantity low-cost roadway assets – such as traffic signs, traffic lights, pavement markings, and guardrails – are critical elements in the operation of transportation infrastructure systems. These assets require preventive, restorative, or replacement work activities to preserve their functionality in an accepted level of service. Nevertheless, in recent decades, the significant expansion in size and complexity of infrastructure networks have posed several new problems on how these assets can be monitored and maintained in a timely fashion. The fast pace of deterioration and the limited funding available have motivated the Departments of Transportation (DOTs) to consider prioritizing roadway assets based on their existing conditions.

Addressing these challenging conditions requires comprehensive, accurate, and frequently updated inventories on the condition of all assets. The key elements toward the development of an asset management program that is capable of producing such inventories are 1) inexpensive and continuous data collection and 2) methods that can further analyze the collected data for condition assessment purposes [1]. The DOT practitioners can then leverage these assessments for maintenance and replacement planning purposes and ultimately improve the condition of the overall transportation systems.

To minimize challenges in data collection, over the past few years, the DOTs have pro-actively looked into road inventory data collection techniques. These techniques involve videotaping road assets – using inspection vehicles equipped with three to five frontal high-resolution cameras – on a truly massive scale. GPS, Inertia Measurement Unit (IMU), and Distance Measurement Indicator (DMI) are often used to provide accurate positional data for these visual sensing systems [2,3].

The level of detail and accuracy required for a high-quantity road asset data collection to document locations, physical attributes, and existing conditions, primarily depends on the intended use of the data [4]. The process of condition assessment using massive visual datasets that the DOTs are collecting today involves reviewing all videos, manually detecting and localizing each asset in relevant video frames, corresponding them to prior assessments (if such database exists), and then performing manual condition assessment based on visual observations [5]. In todays practice, the first few steps are the main bottlenecks of the process. Instead of manually detecting and localizing assets within video frames and matching them to prior assessments which according to our verbal conversations with experts from Virginia and Illinois

* Corresponding author. Tel.: +1 540 235 6474.
E-mail addresses: balali2@illinois.edu (V. Balali), mgolpar@illinois.edu (M. Golparvar-Fard).
[1] Tel: +1 217 300 5226.

http://dx.doi.org/10.1016/j.autcon.2014.09.007
0926-5805/

Departments of Transportation can take up to 60–70% of their time ideally the experts would only spend their time on the more value adding tasks of performing condition assessment on already detected assets and decide on how existing conditions can be improved. Due to high costs associated with the reviews, the number of inspection cycles is very limited [6,7] e.g. a survey cycle of one year duration for critical roadways. This creates negligence for all other local and regional roads which are also frequently used by commuters. The high-volume of the data that needs to be analyzed manually and the subjectivity of the current inspection process even for these selected critical roadways have an undoubted impact on the quality of the analysis [8,9]. Hence, many critical decisions may be made based on inaccurate or incomplete information, which may ultimately affect the asset maintenance and rehabilitation process. Instead of introducing a new method for data collection, this paper leverages existing and already available video frames collected by the DOTs. These videos have high qualities and in particular high spatial resolution making them ideal for computer vision method. Thus, in this paper, we propose a new solution that facilitates the processing of these existing videos. Such system has potential to minimize the need for detection and identifying asset in each video frame, and allows the expert to focus on the more important task of condition assessment.

Automating the analysis of massive visual datasets for detecting, localizing, and analyzing condition of road assets is a challenging research problem. These videos depict large number of similar assets from different camera locations and viewpoints, and have wide variability in terms of illumination conditions and video resolution/quality. Another challenge is the intra-class variability in the visual appearance of the road assets. During the data collection, occlusions are also frequent, and asset positions and orientations may vary [10].

As a first step, research over the past few years has primarily focused on techniques than can automatically detect single type of assets from video streams [11–13]. These methods are mainly introduced and useful for detecting traffic signs; nevertheless, their application for assessing condition of other assets such as guardrails, light poles, and the supporting structures of traffic signs is challenging. Particularly, the 3D pose of assets such as guardrails makes it very difficult to apply template-based computer vision techniques for their detection and localization.

To address current limitations, this paper presents a novel video-based segmentation method which can easily and efficiently segment and recognize these roadway assets from video streams. The method requires no training which makes it scalable to datasets with thousands of video frames and potentially so many categories of asset labels. The proposed nonparametric approach parses video frames and labels image region with their categories of roadway assets. We also present a new benchmark of incomplete and noisy point clouds assembled from a variety of architectural/construction scenes, together with their human-generated segmentations. This dataset provides a benchmark on how expert modelers decompose a point cloud into functional parts, which we treat as "ground truth". Given this dataset, we compute metrics that measure how well the computer-generated segmentations match the human-generated ones. We also compare the performance of our method to the Semantic Texton Forest (STF) method.

The contribution of the paper is the video-based parsing and segmentation methods that can leverage motion cues and temporal consistency to improve the performance of 3D roadway asset recognition. We present metrics and perform quantitative comparisons of our method with the human-generated segmentations and provide a publicly available dataset for future analysis and comparison of video-based asset segmentations: http://raamac.cee.illinois.edu/aca.

2. Related work

In the following, we first briefly review the state-of-the-art 2D asset detection method. Next, recent computer vision methods for segmentation of images and video frames are discussed in more detail.

2.1. 2D road asset detection

The state-of-the-art methods that leverage computer vision algorithms are primarily focused on detecting one type of asset. Ruta et al. [14,15] and Ballerini et al. [16] present algorithms for detecting particular types of traffic signs based on their shape: rectangle vs. triangle. Others focused on only detecting stop and/or speed limit signs [11,13, 17,18]. A more recent line of work such as [19] presents methods for 2D recognition and 3D localization of traffic signs. These methods are primarily applicable to the detection of traffic signs from 2D video streams and are not directly applicable for segmentation and classification of other types of assets including guardrails and light poles. One key idea to address such limitation is to look into image segmentation techniques and partition a video frame into regions that represent a certain type of asset categories [20]. In the following, we present the state-of-the-art segmentation methods from computer vision domain that can segment an image or a video frame into asset categories.

2.2. Image segmentation methods

In computer vision, image segmentation is the process of partitioning an image into multiple salient image regions in which each region can correspond to individual surfaces, objects, or natural parts of objects. To form distinct regions, these methods have conventionally focused on labeling individual pixels with an object/surface category. Shotton et al. [21] method is among the most dominantly used methods. Their work proposes a segmentation method based on bag of semantic textons to group decision trees that can act directly on image pixels. Both textons and priors as features are used to give coherent semantic segmentation and label each pixel. The main drawback is that training generative and discriminative learning models in Semantic Texton Forest method and other segmentation algorithms which operate at the pixel level [21–23] that these methods are fully supervised. This requires providing a fully labeled ground-truth dataset for training purposes. The process of training can take days and must be repeated if new asset categories are added to the dataset. Processing a test image is also quite slow as it involves steps on detecting candidates over an image, performing graphical model inference, or searching over multiple segmentations.

Over the past few years, researches have focused on nonparametric and data-driven approaches that do not require significant training [24, 25]. For each new test image, these methods retrieve the most similar training images and transfer the desired information from the training images to the query image for labeling. Liu et al. [24] proposed a non-parametric label transfer method based on estimating a dense deformation field between images using Scale Invariant Feature Transform (SIFT) flows. SIFT is an algorithm that detects and describes feature points of an image. SIFT is a robust detection and description technique which can handle changes in viewpoint and illuminations (day vs. night) and is fast and efficient enough to run in real-time. The main challenge here is the complex and expensive optimization problem associated with finding the SIFT flow. Moreover, the formulation of scene matching in terms of estimating a dense per-pixel flow field is not necessarily in accord with the intuitive understanding of scenes as a collection of discrete objects based on spatial support and asset category. To address such limitation, Tighe and Lazebnik [25] recently proposed a non-parametric solution to image parsing that is straightforward and efficient. Their proposed method relies only on operations that can easily scale to very large collections of images and sets of labels.

The more fundamental question of whether motion and 3D structure can be used to accurately segment video frames and recognize the object categories is addressed by Brostow et al. [26]. Existing video parsing approaches [26,27] use structure-from-motion techniques to obtain either sparse point clouds or dense depth maps and extract geometry-based features that can be combined with appearance-based features or used on their own to achieve greater accuracy. Their