2017 IEEE Intelligent Vehicles Symposium (IV)
June 11-14, 2017, Redondo Beach, CA, USA

Monocular Localization within Sparse Voxel Maps

David Wong[1], Yasutomo Kawanishi[1], Daisuke Deguchi[2], Ichiro Ide[1], Hiroshi Murase[1]

Abstract— **We introduce a method that uses a single camera to localize a vehicle within a pre-constructed map consisting of a voxel occupancy grid and road-line marker positions. Sophisticated mapping hardware is capable of creating high-accuracy 3D maps of road environments, but localizing a vehicle within such maps is one of the challenges at the forefront of automated driving. A solution which is robust to dynamic environments, while using only inexpensive sensors, is a difficult problem. In addition, maps that enable precise localization consume a lot of data which is impractical for the expansive environments encountered in real-world road networks. We show how using the area of edge regions shared between rendered views of a compact voxel map and in-vehicle camera images can be coupled with non-linear optimization methods to determine the camera position and pose.**

I. Introduction

Vehicle self-localization is an increasingly important area of research, as we move towards fully autonomous driving systems. There are many approaches that aim to overcome the problem of ego-localization within a given map, each with their own sensor requirements and mapping techniques. For automotive localization, Global Positioning System (GPS) receivers are commonplace in navigation products. Aside from very expensive Real-Time Kinematic (RTK) systems, consumer GPS are not particularly accurate and can suffer from large errors when satellite signals are blocked or reflected off buildings and road infrastructure.

For more reliable localization, there are many combinations of sensors and mapping methodologies that can be applied. Important considerations for automotive localization systems include the cost and simplicity of the sensors used in localization, and the size and complexity of the required map. A single forward-facing camera provides an obvious choice as a readily available sensor, but creating a compact map which allows reliable visual localization is a challenge. Traditional methods include matching features extracted from images to create a feature-based map [1], [2], and image databases that use direct visual similarity to determine the camera's location [3], [4]. The increasing availability of inexpensive parallel processing power in GPUs has led to a number of direct localization techniques that render a dense 3D map for direct appearance comparison with input camera images [5]–[7]. In the field of medical imaging, a similar concept is used for the pose estimation of flexible endoscopes in surgical navigation [8], where renders of 3D scans are used for registration of endoscope images. Direct localization methods generally require a high level of detail in the rendering process, with dense 3D models leading to large database sizes.

Our approach employs direct localization techniques, but we base our solution around a compact map which contains much less information than most of the state-of-the-art methods. We use a map created by HERE [9] which contains a series of points representing corners of voxels that make up a coarse occupancy grid. The map provides no color or texture information, so common techniques that use joint image entropies can not be easily applied to the image registration problem. Instead, we apply an image gradient-based objective function that is minimized where mutual edges exist between rendered and real camera images. We then use a non-linear Levenberg-Marquardt [10] optimization to determine the camera pose within the map. The proposed system does not use lighting in the map rendering, and only mutual edges are included in the cost function, allowing camera localization which is robust to the lighting changes and occlusions typically found in traffic environments. Our main contributions can be summarized as follows:

1) We use a compact, texture-less sparse voxel map for rendering map scenes in order to perform localization with a monocular camera.
2) We employ a novel objective function that measures scene similarity based on the area of overlapping edge regions, formulated in a way that enables image comparison across a change in image modality. We also show how the partial derivatives of this function can be derived, allowing alignment of rendered voxel map images to real camera images within a least squares optimization framework.

This paper is organized as follows. In Section II we discuss research that is related to our approach. In Section III we briefly summarize the data and present the problem formulation. We outline our proposed method in Section IV and present the corresponding results in Section V. Following a discussion of the methods and results in Section VI, we conclude the paper in Section VII.

II. Related Work

Vision-based automotive localization research is often approached in a similar way to Simultaneous Localization and Mapping (SLAM) [11]. Many popular SLAM methods use repeatable feature points and descriptors, usually based on the original Scale Invariant Feature Transform (SIFT) [12]. Gradient histograms of local patches at extracted feature points are used to create descriptors that can be repeatably

[1]David Wong, Yasutomo Kawanishi, Ichiro Ide, and Hiroshi Murase are with the Graduate School of Information Science, Nagoya University, Japan davidw@murase.m.is.nagoya-u.ac.jp, {kawanishi, ide, murase}@is.nagoya-u.ac.jp

[2]Daisuke Deguchi is with the Information Strategy Office, Nagoya University, Japan ddeguchi@nagoya-u.jp