ABSTRACT

Shyam Sundar Kannan, Purdue University, December 2024. Monocular Camera-based Localization and Mapping for Autonomous Mobility. Major Professor: Byung-Cheol Min.

Visual localization is a crucial component for autonomous vehicles and robots, enabling them to navigate effectively by interpreting visual cues from their surroundings. In visual localization, the agent estimates its six degrees of freedom camera pose using images captured by onboard cameras. However, the operating environment of the agent can undergo various changes, such as variations in illumination, time of day, seasonal shifts, and structural modifications, all of which can significantly affect the performance of vision-based localization systems. To ensure robust localization in dynamic conditions, it is vital to develop methods that can adapt to these variations.

This dissertation presents a suite of methods designed to enhance the robustness and accuracy of visual localization for autonomous agents, addressing the challenges posed by environmental changes. First, we introduce a visual place recognition system that aids the autonomous agent in identifying its location within a large-scale map by retrieving a reference image closely matching the query image captured by the camera. This system employs a vision transformer to extract both global and patch-level descriptors from the images. Global descriptors, which are compact vectors devoid of geometric details, facilitate the rapid retrieval of candidate images from the reference dataset. Patch-level descriptors, derived from the patch tokens of the transformer, are subsequently used for geometric verification, re-ranking the candidate images to pinpoint the reference image that most closely matches the query.

Building on place recognition, we present a method for pose refinement and relocalization that integrates the environment's 3D point cloud with the set of reference images. The closest image retrieved in the initial place recognition step provides a coarse pose estimate of the query image, which is then refined to compute a precise six degrees of freedom pose. This refinement process involves extracting features from the query image and the closest reference image and then regressing these features using a transformer-based network that estimates the pose of the query image. The features are appended with 2D and 3D positional embeddings that are calculated based on the camera parameters and the 3D point cloud of the environment. These embeddings add spatial awareness to the regression model, hence enhancing the accuracy of the pose estimation. The resulting refined pose can serve as a robust initialization for various localization frameworks or be used for localization on the go.

Recognizing that the operating environment may undergo permanent changes, such as structural modifications that can render existing reference maps outdated, we also introduce ZeroCD – a zero-shot visual change detection framework. ZeroCD identifies and localizes changes by comparing current images with historical images from the same locality on the map, leveraging foundational vision models to operate without extensive annotated training data. It accurately detects changes and classifies them as temporary or permanent, enabling timely and informed updates to reference maps. This capability is essential for maintaining the accuracy and robustness of visual localization systems over time, particularly in dynamic environments.

Collectively, the contributions of this dissertation in place recognition, pose refinement, and change detection advance the state of visual localization, providing a comprehensive and adaptable framework that supports the evolving requirements of autonomous mobility. By enhancing the accuracy, robustness, and adaptability of visual localization, these methods contribute significantly to the development and deployment of fully autonomous systems capable of navigating complex and changing environments with high reliability.