Visual Localization for Autonomous Driving: Mapping the Accurate Location in the City Maze

Dongfang Liu, Yiming Cui, Xiaolei Guo, Wei Ding, Baijian Yang, Yingjie Chen

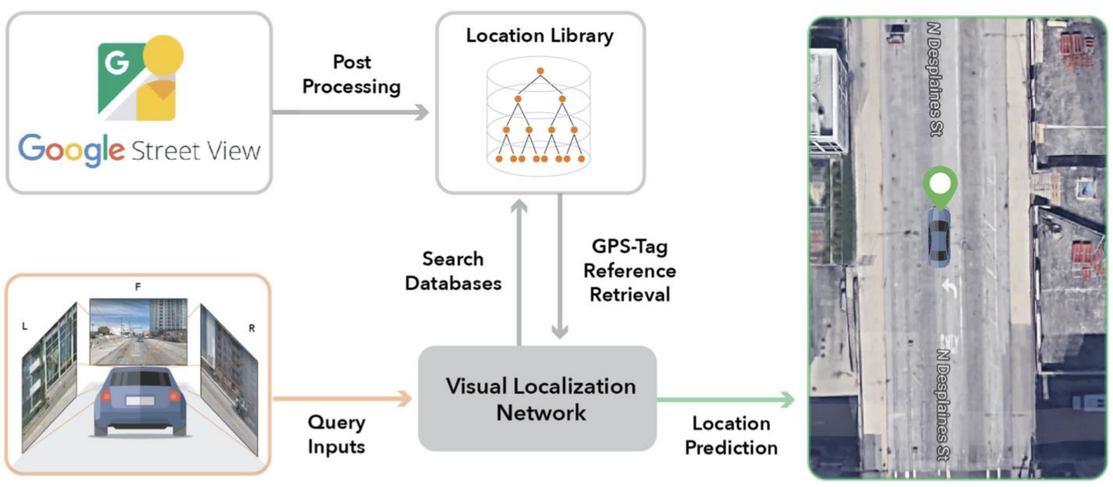
PURDUE UNIVERSITY

Introduction

Accurate localization is a foundational capacity, required for autonomous vehicles to accomplish other tasks such as navigation or path planning. It is a common practice for vehicles to use GPS to acquire location information. However, the application of GPS can result in severe challenges when vehicles run within the inner city where different kinds of structures may shadow the GPS signal and lead to inaccurate location results. To address the localization challenges of urban settings, we propose a novel feature voting technique for visual localization. Different from the conventional front-view-based method, our approach employs views from three directions (front, left, and right) and thus significantly improves the robustness of location prediction. In our work, we craft the proposed feature voting method into three state-of-the-art visual localization networks and modify their architectures properly so that they can be applied for vehicular operation. Extensive field test results indicate that our approach can predict location robustly even in challenging inner-city settings. Our research sheds light on using the visual localization approach to help autonomous vehicles to find accurate location information in a city maze, within a desirable time constraint.

Contributions

- ❑ We design a pipeline to implement a three-directional-view-based visual localization system for vehicles.
- ❑ We implement an automated pipeline to collect, annotate, and manage GPS-tag data for visual localization
- ❑ We leverage the cheap data source from Google Street View which is easy to access.
- ❑ Our method reduces the cost of GPS-tag data collection, annotation, and management.

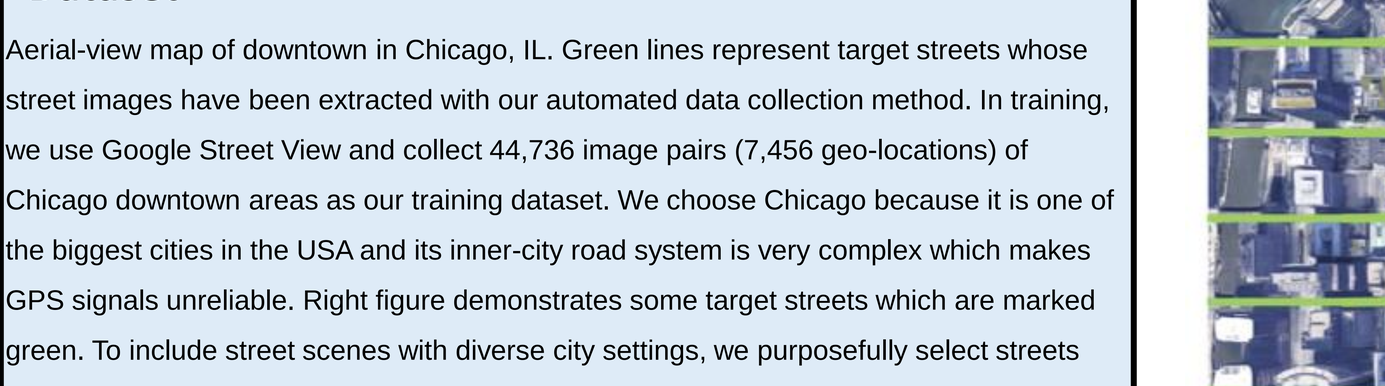


Framework

The framework of the proposed method is presented on the left. We use Google Street View to collect front, left and right view images along the driving trajectory. The collected images are post-processed to construct a location library. In inference, the visual localization network, using the proposed feature voting method, takes the query inputs and searches the location library by matching the local features of the GPS-tagged reference images. The GPS tag from the references with the lowest voting cost is retrieved as the location prediction.

Dataset

Aerial-view map of downtown in Chicago, IL. Green lines represent target streets whose street images have been extracted with our automated data collection method. In training, we use Google Street View and collect 44,736 image pairs (7,456 geo-locations) of Chicago downtown areas as our training dataset. We choose Chicago because it is one of the biggest cities in the USA and its inner-city road system is very complex which makes GPS signals unreliable. Right figure demonstrates some target streets which are marked green. To include street scenes with diverse city settings, we purposefully select streets which include dense districts, crossroads, railway bridges, tunnels, and so forth. We employa vehicle with three cameras mounted on the front, left, and right directions to conduct field tests to evaluate the proposed method (as shown in Fig. 6). Our experimental setting endeavors to simulate the platform of the original Google Street View data collection. We drive the vehicle to repeat approximately the same routes as the Google Street View car to collect the testing data. Our testing dataset includes18,149 image sets, which can be converted into 60 ten-second video clips.



(a) (b) (c) (d)

Results

Qualitative examples are shown on the left. The blue arrows and lines indicate the vehicular trajectory, the green icons are the locations predicted by the visual localization network, and the red icons show the GPS coordinates of the vehicle's assumed position. In this figure, we only demonstrate the results from the D2-Net which has the best performance among all the networks implemented in our experiment. Some quantitative results are shown below.

Method		Runtime (f/ms)	mAP 10(m)	mAP 15(m)	mAP 20(m)
D2-Net	**Three Dimensions**	**42.6**	**91.4**	**95.1**	**98.3**
	Front View	38.2	64.9	70..8	77.2
NetVLAD	**Three Dimensions**	**48.5**	**88.8**	**92.6**	**93.6**
	Front View	40.9	47.8	55.5	68.9
SARE	**Three Dimensions**	**40.6**	**90.2**	**93.2**	**94.2**
	Front View	34.8	50.1	59.8	70.3
GPS	-	-	86.9	92.4	94.6

International Conference on Pattern Recognition 2020