Jie Li

Personal Information

Status:	MS Student
Program:	Computer Science and Engineering
School:	Tandon School of Engineering, New York University
Website:	https://www.linkedin.com/in/jie-li-27b04b5b/
RA Period:	From 2016-02 to 2016-12

Biography

I'm a software development engineer II at Amazon. Before that, I was a research assistant in NYU Multimedia and Visual Computing Lab, advised by Professor Yi Fang. I am broadly interested in 3D Computer Vision, Pattern Recognition and Deep Learning.

M.S. Thesis: 3-D Street Navigation

1 Description

This thesis project explores two existing problems of street navigation: the problem of low accuracy in GPS localization, and the issue raised due to a lack of semantic information during navigation. By using a single image captured by the user, we first propose a novel framework for refining the accuracy of GPS localization and estimating the camera orientation. After that, we further back-project the results of the 2-D object detection on RGB images into a pre-constructed 3D voxel space to perform 3-D semantic labeling. The approach proposed in our project is quite different from any other 3-D street navigation and semantic labeling approaches. Our approach can conduct accurate 3-D street navigation and semantic labeling in real-time, with most of the mobile devices available in the market. Our approach takes advantage of the existing Google StreetView dataset and a state-of-the-art convolutional-neural-network object detector. Ex-

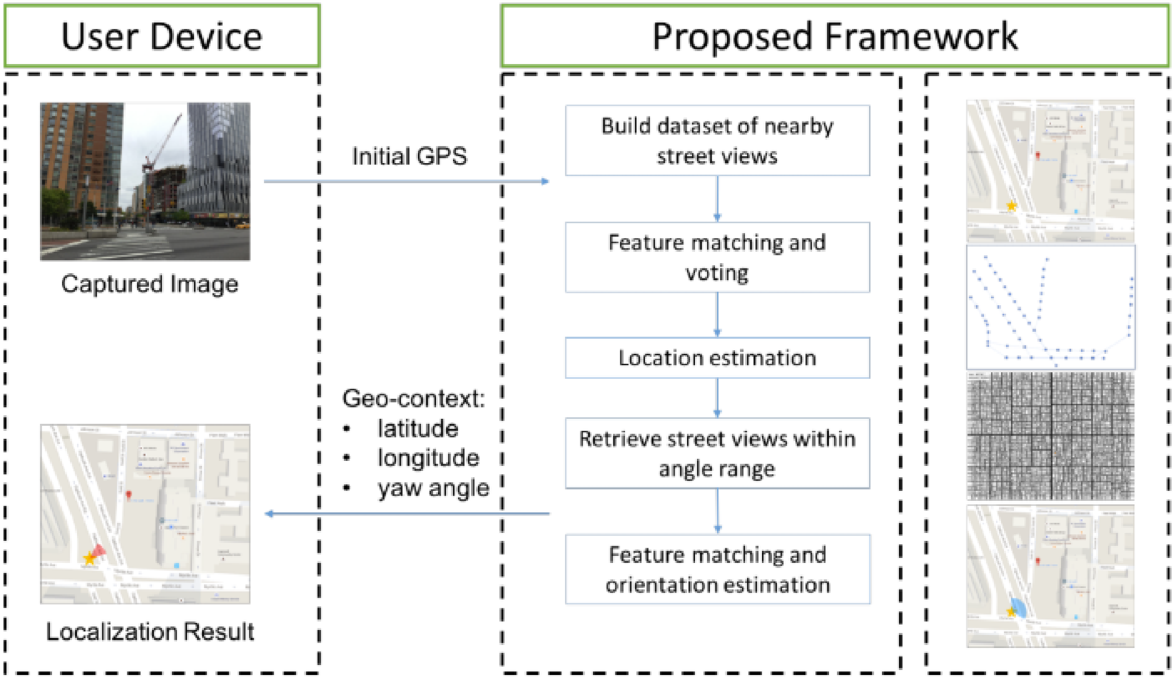

User Device
Proposed Framework
Captured Image
Initial GPS
Build dataset of nearby street views
Feature matching and voting
Location estimation
Retrieve street views within angle range
Feature matching and orientation estimation
Geo-context:
• latitude
• longitude
• yaw angle
Localization Result

Figure 1: 3-D street navigation pipeline.

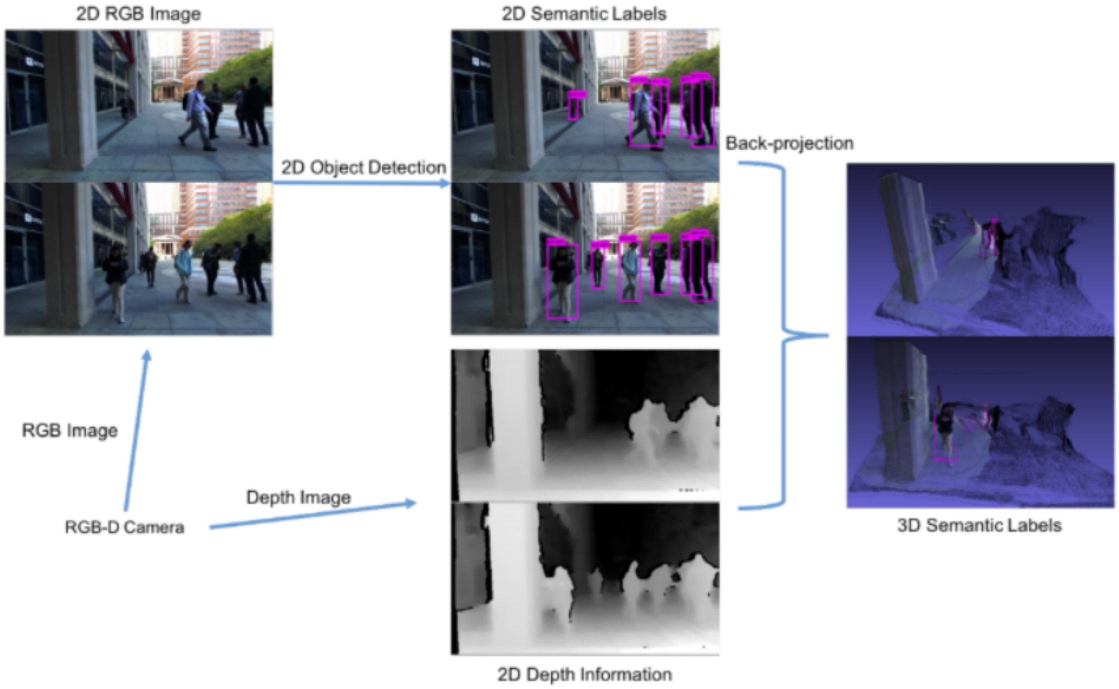


Figure 2: Semantic labeling pipeline.

periments show that our framework significantly improves the accuracy of GPS localization and is capable of providing semantic labels in the 3-D domain at real-time.

2 Method

In this project, the problems of traditional GPS localization and navigation methods are addressed by an image-based localization approach. There are two main components in the pipeline of our proposed approach: a 3-D street localization component to tackle the problem of low GPS accuracy, which is based on Google StreetView, and a 3-D semantic labeling component to take on the problem of 2-D street navigation, based on the depth images. As displayed in Figure.1, the input is an image captured by the user's device and the output is the associated geo-context of the geo-position and yaw angle. After successfully tackling the task of geo-localization, semantic understanding in 3-D scenes is another important topic for the users to identify the objects in surroundings. It is very common that the newly-released mobile devices are equipped with dual cameras. In this case, the dual cameras enable the user to capture depth images from a mobile device. Based on this assumption, we explore a real-time seman-

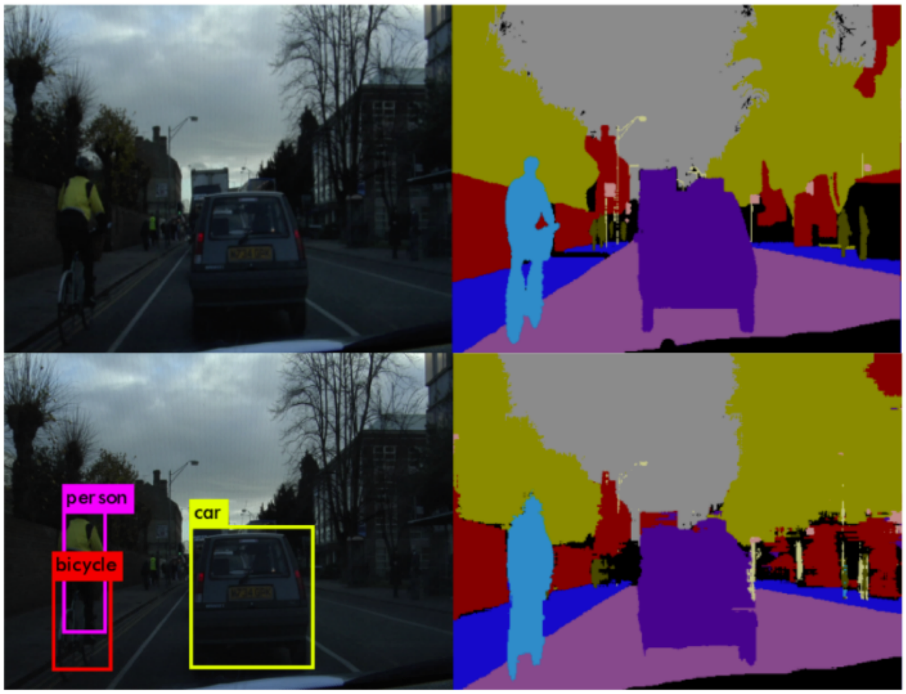
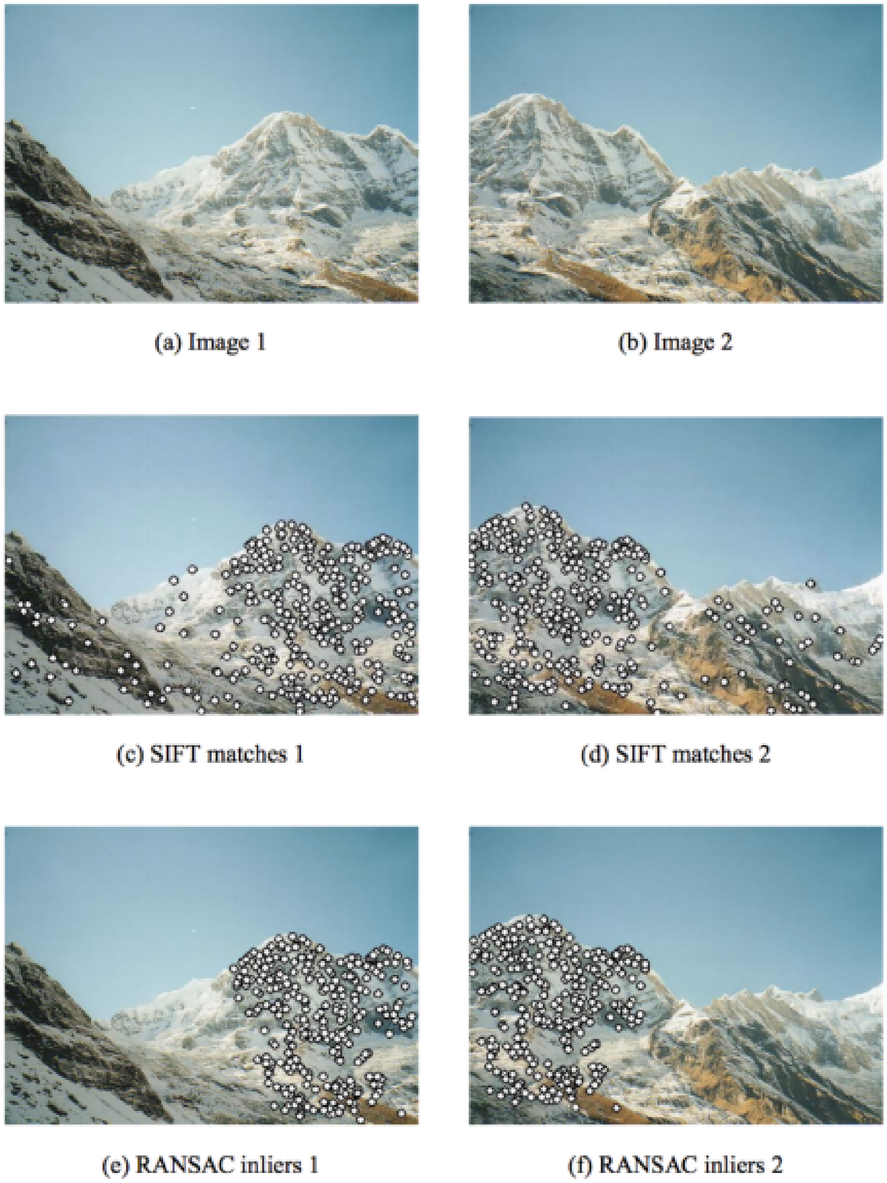


Figure 3: Comparison between YOLO and ReNet on CamVid dataset.

tic object detection approach in 3-D scenes and address the problem of semantic labeling as a cross domain problem from a 2-D image to a 3-D space. Figure.2 shows the 3-D semantic labeling pipeline., The input to the system is a set of RGB-D frames from scanning. Each of these frames has known camera pose.

3 Results

In this section, we first present the experimental outputs. Based on the geolocalization result, the experimental results from our system and the performance comparison between several 2-D object detectors are provided. We will discuss the problems met in experiments and the outcomes achieved in our 3-D street navigation and 3-D semantic labeling components. Figure.4 shows a comparison between image registration with and without RANSAC outlier removal. Images in the first row are the original images to be matched. The second row shows SIFT matches without RANSAC. The last row implies the RANSAC outlier removal result. Figure.3 shows the detection results of YOLO and ReNet on a street frame from CamVid dataset. The two images in the first row are the raw data and the ground truth of semantic segmentation. The left image and right image in the second row are the detection results of YOLO and ReNet respectively. We can see that both YOLO and ReNet succeeded in the detection of the person and the car in the scene. While YOLO can separate the bicycle from the person, ReNet considered these two objects as a single one. On the other hand, ReNet outperformed YOLO in the segmentation of background scenes.



(a) Image 1

(b) Image 2

(c) SIFT matches 1

(d) SIFT matches 2

(e) RANSAC inliers 1

(f) RANSAC inliers 2

Figure 4: Images and SIFt Features.