Knowing Where I Am: Exploiting Multi-Task Learning for Multi-View Indoor Image-based Localization

Guoyu Lu[1]
luguoyu@udel.edu
Yan Yan[2]
yan@disi.unitn.it
Nicu Sebe[2]
sebe@disi.unitn.it
Chandra Kambhamettu[1]
chandrak@udel.edu

[1] Video/Image Modeling and Synthesis Lab
University of Delaware

[2] Department of Information Engineering and Computer Science
University of Trento

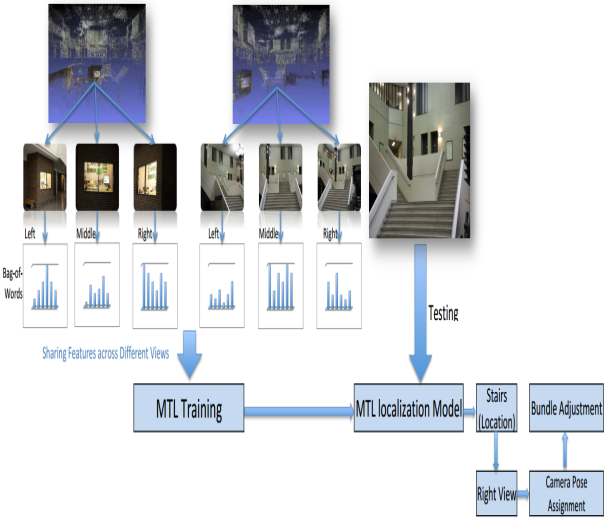

Figure 1: Multi-view image-based localization system

Indoor localization systems are applied to navigate people in large and complex indoor environment, such as shopping malls and museums where auxiliary information is necessary to help visitors localize themselves. In some urgent situation, like boarding an airplane and finding the emergency room in a hospital, providing accurate and timely location information is essential for travelers to catch planes and wounded people to get prompt medical assistance. The majority of the current indoor localization methods is based on WiFi and pre-deployed beacons. These methods usually require additional equipment to perform the localization task and the accuracy depends on the distribution of beacons and cellular stations in a large extent. Meanwhile, the WiFi and beacon based methods are lack of the orientation information, which is essential for navigation. GPS is quite successful in outdoor navigation. However, in indoor buildings with roofs and walls, weak GPS signals result in unreliable navigation information. Even in an outdoor large building area, GPS signals from satellite are attenuated by walls.

Image based localization has been mainly applied in outdoor environments in the past to overcome the weak GPS signal problem among large buildings. This method has been introduced to indoor environments in recent time. The main idea is to linearly search the image database consisting of indoor building images and find the best matched image. With the development of Structure-from-Motion (SfM) reconstruction techniques, 3D models are used for localization. Users can easily capture a 2D image with their mobile phone and register the 2D image with the 3D model to get the location information. In this process, features extracted from the 2D images are utilized to match against the features in the SfM 3D model; camera pose can be calculated based on the matching descriptors, providing users the location and orientation information. As the SfM technique does not require the cameras to be calibrated, the related images are easier to obtain, which makes the large scale reconstruction and 3D model based localization possible. Obtaining the location information is only half of the job. A map with the location information can help better perform the navigation task. With this purpose, a 3D model is suitable for localization purposes that facilitate users to understand the 3D building structure and schedule a visiting plan. However, a SfM model for localization usually contains millions of descriptors. Searching the correspondences within this scope is extremely time-consuming. Although k-d trees and visual word methods are applied to accelerate the corresponding search process, the reduced search scope may potentially add incorrect correspondences between 2D features and 3D points.

In this paper, we propose multi-view image based localization, which is a framework based on multi-task learning (MTL). MTL attempts to improve the performance of several specific tasks based on the shared common properties. Current research shows that it is beneficial to learn the tasks simultaneously instead of learning a single task separately when the tasks exhibit commonalities. During the learning process, the shared information across different tasks is extracted to simultaneously learn the multi-related tasks. With the purpose of guiding users with the location and orientation information, we divide the physical view direction into several regions. It is expected that images of the same object captured from different view directions contain similarities with regards to appearance, as well as differences due to the viewing perspectives.

Multi-view image based localization aims to learn the relationship of interior architecture appearance across different viewing directions. Ideally, the tasks within the same group should share the similar features while features extracted from tasks in different groups are expected to be different. Following this idea, images captured from the same direction are classified into one task, including same and different location images. The images captured from the same location across different camera angles are treated as the same group. We learn a multi-view regression model based on the correlated tasks scattering in different groups. During the testing phase, the query image retrieves the most relevant group for achieving the location information. Meanwhile, our MTL regression model assigns a direction to the query image based on multiple tasks for the orientation purpose. As we perform SfM reconstruction prior to the multiple view localization phases, every image used for SfM reconstruction is associated with a camera pose. The camera pose of the most corrected image within the same task, and the same group is assigned to the query image. We further apply bundle adjustment to the query image to refine the assigned camera pose. In this way, we can take benefits from localization methods both based on 2D image and 3D model. The whole multi-view image-based localization framework is illustrated in Figure 1.

To summarize, the contributions of this paper are the following: (i) To our knowledge, this work is the first to address the problem of indoor image-based localization from multi-view settings. (ii) We are the first to propose the multi-task learning approach for multi-view indoor image-based localization. (iii) Both the orientation of the image and the location information can be obtained by exploiting multi-task learning.

Making use of the multi-task learning method, we develop a multi-view image based localization system. By separating the view directions into 3 different partitions as tasks, we simultaneously learn the relationship among the tasks, which can improve the prediction accuracy of each view orientation. The learned multi-view regression model can accurately retrieve the location information. After learning the model, our multi-view system can retrieve the location and view orientation information by computing a dot product to assign a correlation score, avoiding large scale correspondences search. Leveraging the 3D localization system, we assign the camera pose of the nearest neighbor image of the same orientation and location used for SfM reconstruction to the query image, with further refinement using bundle adjustment. Embedding our multi-view method into the 3D localization system helps us better achieve the localization information in a 3D map.