A Joint Framework for Object Recognition: Categorization, Instance Recognition and Pose Estimation

Tarek El-Gaaly
Rutgers University
10/16/2015 at 02:00 pm
CBIM 22

Abstract

Visual object recognition is a challenging problem with a wide range of real-life applications. The difficulty of this problem is due to variations in appearance and shape among objects within the same category, as well as varying viewing conditions, such as viewpoint, scale, illumination, occlusion in cluttered environments and articulation in multi-part objects. In addition, beyond the visual spectrum, depth sensors suffer from noise that inhibit the ability to perform robust object recognition. Under object recognition lie three subproblems that are each quite challenging: category recognition, instance recognition, and pose estimation. Impressive work has been done in the last decade on developing computer vision systems for generic object recognition. Despite this, the problems of multi-view recognition and recognition from 3D data remain some of the most fundamental challenges in computer vision.

In this dissertation we focus on the goal of jointly solving the three subproblems of object recognition from multi-view images and over multiple modalities; from images to depthmaps and 3D point clouds. There are two main parts to this work. These two parts explore different representation spaces for solving joint object recognition. The first part delves into representation spaces that allow manifold analysis to solve the three sub-problems of object recognition from multi-view images. The second part explores a representation for 3D point clouds of multi-part objects that is inherently pose-invariant and provides valuable shape information for object categorization.

Multi-view images of the same object lie on an intrinsic low-dimensional manifold in descriptor spaces. These object-view manifolds of different objects share the same topology despite being geometrically different. For this reason object-view manifolds can be parameterized by their topological mappings to one unified manifold, e.g., a viewing circle in the case of a fixed camera.
capturing an object on a turn-table. We explore ways of parameterizing these representational mappings in order to solve for category, instance and pose. We also extensively analyze the layers of a Convolutional Neural Network (CNN) to see their effect on object-view manifolds and propose multi-task CNN models to jointly solve categorization and pose estimation.

The second part of this dissertation explores building a representation for categorization and pose estimation of more complex multi-part objects purely by shape information, in the form of 3D point clouds. Shape is a powerful cue and can efficiently augment vision for object recognition. We build a model to decompose 3D point clouds of objects into parts and inherently encode pose information using skeletal representations. The resulting representation is robust to pose variation, deformation and articulations and provides a robust representation for multi-part joint object categorization and pose estimation.