Reconstructing 3D Structure from Amazon Product Images

Jason van der Merwe
jasonvdm@stanford.edu

Andrew Giel
agiel@stanford.edu

Bridge Eimon
beimon@stanford.edu

CS 231a Stanford University
December 8, 2013

Abstract

Constructing 3D structure from two dimensional images usually requires many images, covering a variety of angles of the object you are trying to reconstruct. However, in this case, we only had a few images, usually only two.

1. Introduction

This project, reconstructing 3D models from Amazon product images, is the first step of a larger goal to build a database of 3D objects by crawling the web. This database can then be used to assist in many other object detection tasks, such as identifying objects in your home or from a cell phone camera. One important component of making such a system is to build 3D object models from multiple views of web images. SFM is generally performed using a number of images carefully taken of the same object. Can we use SfM from product images taken by different users? Amazon often has a collection of images for each object uploaded by different users. These views were all taken with different cameras in different lighting / backgrounds, which adds a challenge. Can we combine these images to create 3D object models?

2.1 Previous Work

Recreating the structure of an object from a set of images has been explored by many instances of previous research, even in a completely unsupervised and automatic fashion.

In fact, the concept of taking images available online of the same object or scene and then reconstructing the three dimensional properties from these images has been achieved by previous research. In a paper entitled 'Photo Tourism: Exploring Photo Collections in 3D', Noah Snavely presents a system that retrieves images of famous landmarks and then infers the structure of the landmarks in three dimensions as well as the pose and location of the cameras which captured the images. After establishing correspondences, Navely utlized the Levenberg-Marquardt algorithm to solve the least-squares problem of sparse bundle adjustment. Once this structure known, Snavely and his team were able to rerender the scene as well as the position of the cameras.

A key portion of the reconstruction process, as mentioned above, is establishing correspondences between the images. This can be achieved via expensive manual annotation or by solving the longstanding correspondence problem. David Lowe's work on the Scale-Invariant Feature Transform (SIFT) helped to address this cardinal problem in the vision commmunity. By implementing a difference-of-Gaussian function to identify points of interest in the images, Lowe was

able to then localize and establish their dominant orientation. Once the keypoints were defined and desciptors established, Lowe utilized best-bin-first search in order to find the best match from any one keypoint of an image to a keypoint from another image, using Euclidean distance as the ranking metric. In order to determine the probability of a match, Lowe used the ratio of the distance to the best match to the distance of the second best match, eliminating matches with a distance ratio over a certain threshold. Combining this methodology with a model-generating procedure such as Random Sample Consensus (RANSAC) has proven to be effective in estimating the correspondence problem, and was utilized in our implementation as well.

Given a set of correspondences, recovering the structure from motion is a non-trivial problem. Tomasi and Kanade, in their seminal paper, present a method for extracting scene geometry and camera motion from a matrix W derived from the measurements and correspondences. Tomasi and Kanade realized that the points are represented within a low-dimension subspace of W, allowing them to be extracted via Singular Value Decomposition. This method performs best when used on points without occlusions, yet methods are presented to extrapolate measurements.

2.2 Advancements

3.1 Technical Summary

Before we began using Amazon product images, we began with a proof of concept using photos of objects that we took ourselves. Once we took photos of these images in a controlled environment, we began implementing the technical aspects of the project. First, we manually selected correspondence pairs and then computed SFM using the Tomasi-Kanade factorization algorithm. We also computed the fundamental matrix using the normalized 8-point algorithm. After we were happy with these results, we moved to automatically retrieving correspondence pairs. We implemented SIFT, SURF and ORB and compared the results. We ran all matches through RANSAC, experimenting with different thresholds. Plotting these points in a three dimenionsal space showed us the recovered structure of the product.