Dense 3D Mapping with Monocular Vision: 
Bridging the Gap Between Robotics and Computer Vision

Kamil Wnuk
Computer Science Department, Harvey Mudd College
340 E. Foothill Blvd., Claremont, CA 91711
kwnuk@cs.hmc.edu
Advisor: Zachary Dodds

Abstract

Currently, the vast majority of autonomous mapping in robotics relies on direct measurements from costly devices such as sonar, infrared, and laser range finders. Adapting established methodologies from the structure−from−motion (SFM) subfield of computer vision to data commonly available in robotics, we have created a unique toolkit, able to render visually dense 3d maps from odometrically annotated monocular vision. Though computationally intensive, monocular vision provides a low−cost, yet highly capable alternative to currently popular mapping sensors. Future research on application of our work for navigational uses hopes to create inexpensive monocular robotic systems capable of autonomous navigation.

Problem and Motivation

As humans, most of us take the computation involved in our visual perception of the world for granted. Despite efforts in the robotics and computer vision communities, a method of representing visual data to guide an agent around an environment, that is as efficient and effective as human vision, has not yet been developed.

The research we are conducting contributes to the solution of this problem by focusing on map building with a very limited mobile sensor suite. Our newly developed toolkit allows for construction of visually dense 3d maps from data gathered by a platform equipped solely with monocular vision and odometry.

Currently, the vast majority of robotic navigation relies on direct range−to−obstacle measurements from devices such as sonar, infrared, and laser range finders. We believe that dense 3d maps can, themselves, serve as the raw material for algorithmic navigation techniques such as motion planning and localization. Thus a tool for constructing such maps would be beneficial to the robotics community. Using only monocular vision for sensing allows easy adaptation to any mobile platform. Though the computational requirements are high, vision is a much more cost−effective alternative to current navigational sensor suites.

Background and Related Work

As noted above, vision is not usually the tool of choice for roboticists. When vision systems are utilized, the data representation chosen is often sparse. For example, Kosecka and Li, [1], used scale−invariant (SIFT) features and histograms of 2d images to localize in indoor environments. Another such example is a monocular map building system developed by Neira et al., [2]. Their maps were also sparse, based on vertical edges in the environment. A number of other sparse mapping techniques have been developed using different permutations of stereo cameras, laser range finders, and panoramic vision.

The structure−from−motion (SFM) subfield of the computer vision community, on the other hand, has experienced great success creating dense 3d reconstructions from visual data, as demonstrated by Pollefeys et al.’s method of modeling structure using a hand−held video camera, [3].
Our approach bridges the gap between computer vision and robotics by adapting established SFM techniques, [4], for use in traditional robotics domains. Our method takes advantage of robot odometry and the planarity of indoor environments to create visually dense 3d environmental reconstructions using a single camera. Just as our robotic focus differentiates our work from mainstream computer vision, the limited nature of our sensor suite also distinguishes us from sensor–rich world acquisition methods more commonly found in robotics.

**Approach and Uniqueness**

The map building process consists of five principal stages: feature identification and tracking, feature triangulation, surface fitting, estimation of homographies between images and the 3d surface, and mosaicking. This section will give a brief overview of each of these stages, demonstrating the novelty of our approach.

The feature identification and tracking on the input image sequence relies on Stan Birchfield’s KLT tracker [5]. This first processing stage outputs the coordinates of every tracked feature, for each image that it appears in. Each feature is assigned a unique ID upon its first tracking between images, and maintains that ID as long as it continues to be tracked through the sequence of images.

These tracked features’ coordinates are then passed to our Triangulator, along with the camera parameters, acquired via an earlier offline calibration. From odometry the Triangulator computes the camera rotations and translations for each image, followed by a least–squares estimation to obtain the 3d coordinates of each of the features. A critical component of the Triangulator is that it allows for feature IDs to be propagated to the 3d points, thus maintaining a direct correspondence with the original images.

Once the coordinates of the features in 3d space are known, a RANSAC [6] algorithm is used to find the optimal set of planes to fit the points. For each plane the algorithm generates 2000 hypotheses and the optimal fit is decided based on a set of heuristics. This subsystem outputs a set of plane normals, a list of feature points to which each plane is fit, and the 2d coordinates of each feature when projected onto its plane.

The final two stages are responsible for mapping visual data onto the estimated structure. For each plane in the structure, a separate texture image is created. Homographies between sequential images and the environment’s planes are found using the results from the tracking phase coupled with data from the plane fitting. These homographies are passed to the mosaicker, which creates a texture image for each plane. To accompany each texture is a file holding the coordinates of all the features appearing in the mosaic and their IDs. This allows the texture to be correctly aligned with the points in 3d space to create a visually dense map of the environment through which the robot has traveled.

To see a data flow diagram explaining the map construction process, please visit http://www.cs.hmc.edu/~kwnuk/wart/architecture.jpg.

**Results and Contributions**

Using our toolkit we were able to successfully map a portion of our lab as well as a nearby hallway, thus validating our approach. To see these results, along with images from various intermediate stages, see http://www.cs.hmc.edu/~kwnuk/wart/results.

A unique ability provided by our maps is the capability to render views of the mapped environment from previously unvisited poses, allowing for the generation of anticipatory visual data. We intend to further build on our current work by utilizing such renderings of maps constructed by our system to develop new monocular–vision–based navigational algorithms.
To promote continued work in applying SFM methodologies to robotics, the full source code (C++ for Windows) of our toolkit is available for free use at http://www.cs.hmc.edu/~kwnuk/wart. We feel that further work in the fusion of these two fields will endow the robotics community with highly capable, very low cost robotic platforms, thus making vision–based robotics accessible far beyond research laboratories.

References


