High-Resolution Mapping and Modeling of Multi-Floor Architectural Interiors MIT9904-20

Progress Report: January 1, 2001—June 30, 2001

Seth Teller

Project Overview

The long-term goal of our research is to develop autonomous (robotic) mobile sensors capable of moving around in an architectural environment, observing the environment, and constructing a useful, textured geometric CAD model of the environment. We envision dispatching several such sensors in parallel into a previously unknown environment. The sensors will then collaboratively explore the environment, and the end product will be a useful map, or model.

Our NTT-sponsored efforts focused on a subset of this long-term goal: namely, the ability to localize and fuse a low-resolution omni-directional video stream gathered by a rolling sensor. We have pursued this research on several fronts: developing improved algorithms for processing omni-directional video; developing sensors (in collaboration with Prof. Balakrishnan) for improved camera localization and heading determination indoors.

Progress Through June 2001

In the city scanning project, one fundamental technical obstacle has been the development of robust algorithms to determine 6-DOF pose (position and orientation) for the moving sensor. Our primary technical achievement in that project is the development of fully automated algorithms for exterior calibration (pose recovery) for networks of thousands of images over extended areas spanning hundreds of meters. Our algorithms are accurate to roughly a tenth of a degree of absolute rotation, and five centimeters absolute translation (see http://graphics.lcs.mit.edu/~seth/pubs/pubs.html for relevant publications). These algorithms have been published in the CVPR (Computer Vision and Pattern Recognition) conference in 1999, 2000, and 2001 (two papers have been accepted for publication).

We observe that the data stream gathered by our sensor is similar, in principle, to the data stream gathered by our outdoor sensor, the Argus, in that it contains wide-field-of-view imagery, and that each image is annotated with a rough estimate of the position and attitude of the acquiring camera. We reasoned that many of the techniques applied in the outdoor regime should be applicable to indoor data. However, there are significant differences in the data characteristics which have made extension to indoor operation difficult.

Outdoors, we can reliably detect families of parallel lines from buildings, and exploit GPS for good initial position estimates. Also, the Argus sensor includes a high-resolution camera and mechanical pan-tilt head, enabling the capture of high-resolution spherical mosaics. (A typical Argus mosaic has spatial resolution of about one pixel per milliradian, or 1,000 pixels in a roughly 60-degree field of view.) Our pose estimates are good to within 1-2 pixels rotationally, and 1-5 pixels translationally.

Indoors, the situation is different. GPS is not available, so we have no a priori translation estimates. We can continue to exploit dead reckoning information from odometry, however. Our image sensor, a CycloVision OmniCam, operates at NTSC resolution, about one one-hundredth the resolution of the Argus. (This is about one-tenth the Argus's linear resolution, or about half a degree per pixel.) Whereas outdoors we acquire about one image per minute, with ten-meter baselines between images, indoors we acquire images at 30Hz – a thousand times as fast – and with baselines of only a few centimeters for a slowly moving platform. Finally, whereas outdoors the illumination conditions are uncontrolled and highly variable, indoors we can expect lighting with a lower dynamic range and absolute scale. Indoors, we have found in practice that contrast is often low, which makes edge and feature detection less reliable. All of these factors amount to a significantly different deployment environment indoors, when compared to the outdoor setting.

The spherical imagery from the OmniCam is well suited to our algorithms, which exploit vanishing point and expansion center geometry on the sphere. We have extended our camera models from standard to omni-directional imaging geometry, and processed omni-directional video sequences through our end-to-end processing pipeline. The pipeline is currently a batch process, and far slower than real-time (30Hz); one area of current investigation is achieving a real-time or near-real-time processing capability, so that annotated video imagery can be localized, and a crude model extracted, as the environment is observed rather than later. We have also continued a collaboration with Prof. Michael Black at Brown University on developing dense optical flow algorithms for registering omni-directional video.

Research Plan for the Next Six Months

We are investigating a number of techniques for the acquisition system over the next 6-12 months. First is the continued use of motion factoring to increase robustness in egomotion estimation. In our case, the 6-DOF rigid transformation relating each frame of the sequence can be decoupled into a pure 3-DOF rotation, a pure 2-DOF translation (i.e., a direction up to unknown scale), and a pure 1-DOF absolute scale. In the absence of GPS information, scaling information must be provided by metric odometry, or by fiducial features or objects having known size or separation in the scene. We are evaluating several possibilities for providing metric information, including the use of fiducial position transceivers (Crickets) as described below.

We are also working with Prof. Hari Balakrishnan to develop and exploit sensor-based position and attitude determination capabilities using the Cricket architecture. In a way analogous to GPS, an environment can be instrumented with a set of fixed transceivers, each of which broadcasts its known position and a unique identifier. A mobile receiver can then combine several received signals to infer its own position. We are extending this capability in two ways. First, by adding multiple receivers to a small hand-held device, we can infer the receiver's attitude as well as position. This yields a valuable egomotion estimate to the mobile image sensor; it is absolute, rather than relative, as would be an egomotion estimate derived solely from odometry. This work has been published in MobiCom (Mobile Computing) 2001. In the next 6-12 months, we will further develop the attitude determination capability, for example by making it more robust, and available throughout the building.

We are integrating the Cricket capability into our new NTT research effort, exploitation of existing CAD models of campus buildings for tasks such as resource location and path-finding. In this scenario, models produced by the acquisition system serve as natural embeddings for information about people, devices, and services associated with the space of the model. Thus, a person unfamiliar with the space could be assisted in locating relevant services with an appropriate sensor – analogous to the way a car-based GPS system helps a driver find a desired restaurant or hotel. We plan to associate office inhabitants with the building model, and develop queries to (for example) determine the most efficient route to a particular person's office.

On the sensor side, we will continue to improve our numerical estimates of the camera's internal calibration, and extend our egomotion estimation algorithms to incorporate translation estimation.

Finally, we will continue the development of robust, scaleable spatial data structures and algorithms for handling complexity, in the form of very large numbers of observations and output features. In particular, we use spatial data structures that support inverse range queries, so that (for example) given a region of interest, we can rapidly identify the data elements – image and navigation data – to support 3D reconstruction, model acquisition, and location-based queries in that region. One particular area of interest is super-resolution reconstruction: since we are acquiring frames at 30Hz, we will have hundreds or even thousands of observations of many surface fragments. We will develop methods to combine these numerous, noisy observations to produce a single, high-quality estimate of the geometry and texture of the observed surface fragment.