Vision Based Localization and Environment Modeling for Unmanned Vehicles

Piotr Jasiobedzki, *Member, IEEE;* Stephen Se, *Member, IEEE*

***Abstract*— Safety and operational demands require that operators of unmanned security and defence vehicles be located at safe distances. The capability of creating photorealistic 3D models using on-board sensors on unmanned vehicles will improve the operators' situational awareness.**

Over the last years MDA in Brampton have been developing computer vision technologies for localization and 3D modeling of environments and objects from mobile cameras. The technology relies on real time image acquisition and processing using a combination of dedicated hardware and software. The basic system processes images from stereo cameras and computes the camera motion in 3D (position and orientation). A real-time version has been deployed on a mobile rover and has been used as a visual odometry sub-system for off-road navigation.

The on-going work focuses on augmenting iSM with additional sensors (Infrared and Gamma) and creating multi-modal representations of environments contaminated with Chemical, Biological, Radiological and Nuclear agents.

***Index Terms*—stereo cameras, visual odometry, 3D modeling**

I. Introduction

Safety and operational demands require that operators of unmanned security, defence, and urban search-and-rescue vehicles be located at safe distances. The vehicles may be operating under direct teleoperation or be endowed with certain levels of autonomy. In the teleoperation mode the operator relies on on-board camera images to navigate around obstacles and execute the mission. The vehicle may be equipped with a Global Positioning System (GPS) and inertial sensors that provide the current vehicle location and heading in a geocentric coordinate system. This information may be visualized together with local maps on the operators' consoles to help them in situational awareness.

In a semi-autonomous mode the vehicle may use GPS for localization in the geocentric coordinate system and, by following predefined waypoints, may automatically stay on a path leading to successful execution of the mission. The waypoints are typically defined by the operator in advance or may be computed automatically using global goals, and e.g., local maps and traversability information. On-board sensors can be used for detecting obstacles not present on the map and avoiding them. These obstacles may not be known in advance; due to their small size relative to the map resolution they may not have existed when the map was created or may be of transient nature.

In both cases of the teleoperated and semi-autonomous vehicles the operator relies on prior maps to understand the vehicle surroundings and covered path. The maps may be updated with obstacles detected by the navigational sensors. Video footage from on-board cameras provides additional information. In general, the situational awareness is rather limited, relying heavily on the operator memory of the visited sites and two dimensional (or elevation) maps. There is a need for systems that can process images from on-board cameras and sensors and create photorealistic models of visited locations in near real-time (seconds to minutes from data acquisition). Such models should be calibrated, allowing the operator to perform interactive measurements and plan operations. The models should be registered with geographic coordinate systems and available maps.

This paper describes the development of 3D modeling systems carried out at MDA. Such systems can create calibrated 3D photorealistic models of visited environments and observed objects from images acquired by stereo cameras on-board unmanned vehicles.

II. Approach

An architecture of the 3D modeling systems developed by MDA is shown in Figure 1; detailed descriptions can be found in the following publications [1, 5, 6].

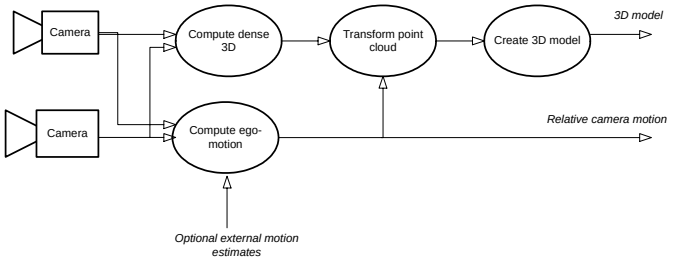

Figure 1 Architecture of the 3D modeling system

Images from two or more stereo cameras are processed in two separate paths:

- Depth estimation
- Ego-motion estimation

The depth estimation module uses the stereo image pairs acquired from synchronized cameras to compute a depth

Piotr Jasiobedzki is with MDA, Space Missions, 9445 Airport Rd., Brampton ON L6S 4J3, Canada, phone 905 790 2800 x 4647, e-mail: piotr.jasiobedzki@mdacorporation.com

Stephen Se is with MDA, Space Missions, 9445 Airport Rd., Brampton ON L6S 4J3, Canada, e-mail: stephen.se@mdacorporation.com

image. The depth estimation assumes fixed and calibrated camera geometry and involves typical stereo image processing steps: image rectification, dense disparity calculation, validation and conversion to 3D using the known camera geometry.

The ego-motion estimation module detects local features of interest (tie-points) in the stereo images and computes their locations in 3D space in each stereo pair. The features are detected and matched using the Scale Invariant Feature Transform (SIFT) algorithm [2, 3], which allows identifying the same features reliably in images taken from different distances and from slightly different viewpoints. Feature matching relies on stereo geometry constraints and similarity of features' appearances. Typically, hundreds of features are detected in each stereo pair.

Features detected in new images are matched with features detected previously and stored in a features database using the Simultaneous Localization and Mapping (SLAM) approach. As the SIFT features have rich descriptions and are largely scale and viewpoint invariant they can be matched with a very small number of outliers. By computing a rigid body transformation that minimizes the overall distance between all matched features, the ego-motion module estimates the camera motion in 6 degrees of freedom. Any outliers are detected then and rejected from the data set; only accepted features are added to the database.

The camera motion may be estimated exclusively from the stereo image sequences as described above or may also be combined with motion estimates provided by other sensors. For example, if the camera is used on-board of a moving vehicle data from wheel odometry or inertial measurement unit may be fused with motion estimated from images. When the camera is mounted on a positioning device (manipulator or pan-and-tilt unit) the telemetry may be provided as an initial motion estimate and refined using vision. In the hand-held mode an orientation sensor may be attached to the camera to provide measurements that are fused with vision based motion.

The dense 3D data computed from individual stereo pairs is mapped into one coordinate system typically associated with the first frame. As there is a significant amount of overlap between the images (which is required to compute the motion), the resulting 3D data contains redundant measurements. These are averaged, noisy measurements are eliminated and the 3D point data is converted into a 3D surface represented as a triangular mesh. The photorealistic appearance of the 3D model is created by applying regions selected from original images as textures on the 3D surfaces [5].

Multiple scans with overlapping regions can be registered automatically by matching SIFT databases obtained for individual scans. This allows for creation of larger models and obtaining full scene coverage [6].

As the stereo cameras are calibrated, the resulting 3D models are also calibrated, which means that the system operator may select points in the model and measure distances, angles, area and volume. If the cameras are mounted on-board a vehicle the estimated camera motion can be used for the vehicle localization and navigation.

III. Implementations

We have developed and tested several prototype systems:

- Hand-held
- Motorized
- Vehicle mounted

A. Hand-held system

The hand-held version of the system, the instant Scene Modeler (iSM) is shown in Figure 2. iSM uses a commercial stereo camera, Bumblebee from PointGrey Research [4], and a laptop computer for data acquisition, processing and model visualization [5].

Figure 2 Hand held instant Scene Modeler (iSM)

iSM captures stereo images at a maximum supported frame rate (7-15 Hz) and stores them in the laptop's memory. The camera motion estimation is based on the processing of stereo images alone and is performed purely in software. The software estimates all 6 degrees of camera motion (translations and rotations). A sequence of approximately 30 seconds' duration is processed in several minutes on a standard laptop. Figure 3 shows one of the stereo images from a sequence and Figure 4 the reconstructed model.

Figure 3 One of the images of a stereo sequence

Figure 4 A model reconstructed from the sequence

B. Motorized

The motorized and tripod mounted version of the system is used when automatic image acquisition and full scene coverage are required. A prototype developed for modeling of underground mines is shown in Figure 5. A custom stereo camera with a longer baseline allows for imaging of objects at longer distances and an integrated camera light provides the necessary illumination. The camera head is mounted on a motorized pan-and-tilt unit and moves to pre-programmed positions, recording images and telemetry.

Figure 5 Motorized version of the system

The motion estimation module uses telemetry as an initial guess for the camera pose--which may be refined through the vision based processing described above. Two views of a reconstructed underground tunnel are shown Figure 6. Multiple scans have been acquired from different positions and registered together using the overlapping sections.

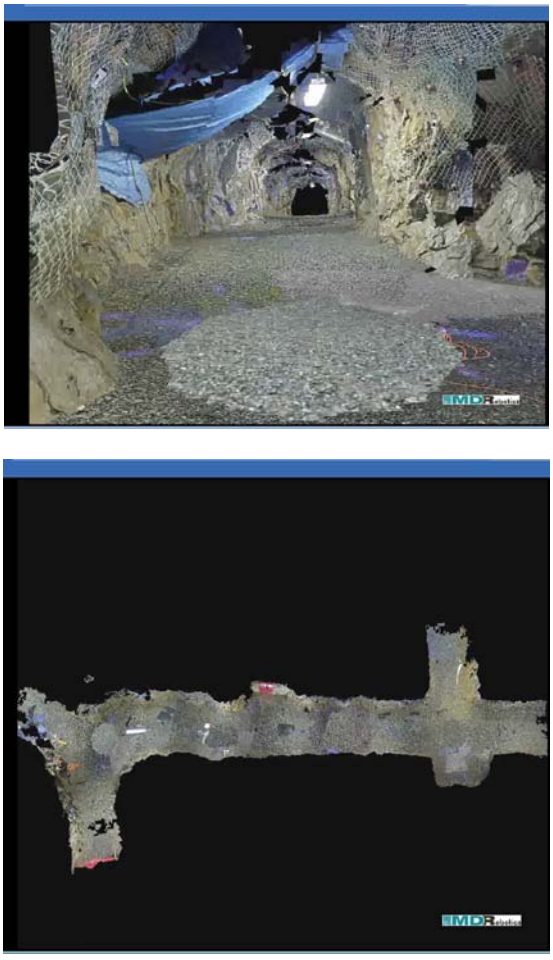

Figure 6 Reconstruction of an underground tunnel (tunnel view and top-down view)

C. Vehicle mounted

The stereo cameras together with the data acquisition / processing computer have been integrated with prototype mobile robots: one of the configurations is shown in Figure 7 [1].

Figure 7 Vehicle mounted stereo cameras and vision system

Stereo images captured from a moving or stationary vehicle are processed to create 3D models using both modes described above. Example images selected from a sequence collected

from a moving rover during tests in a desert are shown in Figure 8.

Figure 8 Images selected from a sequence of 400 images captured from a moving mobile robot

Processing of the approximately 400 stereo pairs resulted in a 3D model shown in Figure 9. Image on the top has been rendered from an approximate initial camera location. The image on the bottom shows the top view of the scene; the red/green/blue lines represent the camera locations for each frame during the image acquisition.

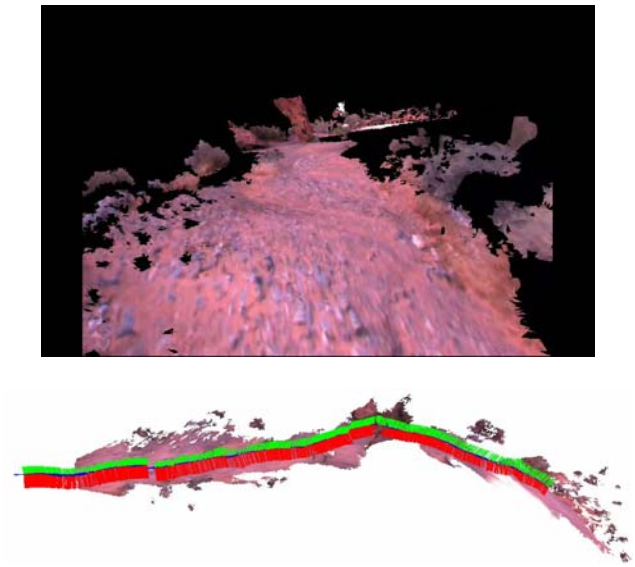

Figure 9 Two views of a 3D model reconstructed using images from a moving robot

The on-board vehicle computer is equipped with a dedicated hardware processor based on Field Programmable Gate Arrays (FPGA) that computes SIFT features at rates of 7 Hz for 1 Mpixel images [1]. This allows for using the camera ego-motion for visual odometry--to estimate the vehicle location in real-time. Unmanned and autonomous vehicles often rely on wheel odometry and inertial sensing to estimate their locations in the absence of GPS signals. Wheel odometry incurs significant errors due to wheel slippage on soft terrain and inertial sensors accumulate error over time. A comparison of visual odometry with wheel odometry and differential GPS for a 120 m test run is shown in Figure 10. It can be seen that the visual odometry (blue) is very close to the differential GPS (green) while the wheel odometry (red) drifts off quickly.

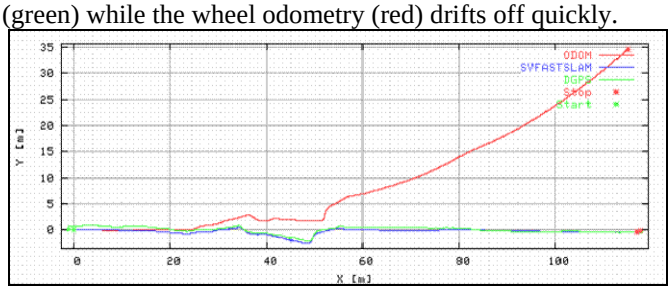

Figure 10 Comparison of errors incurred by wheel and visual odometry

When operating on vehicles the ego-motion estimates are fused with the measurements from wheel odometry and inertial sensors.

IV. Conclusions and On-going Work

This paper presents a brief overview of the research and development on 3D modeling system conducted at MDA. The approach relies on the processing of sequences of images from mobile stereo cameras and estimating camera motion and range data automatically. The estimated camera motion is used to register multiple range data sets and to create photorealistic 3D models. When implemented on dedicated hardware (FPGA) the motion is estimated at rates of several Hz per image allowing it to be used for visual odometry and vehicle localization.

Current work at MDA focuses on adaptation of the 3D modeling technologies for investigating scenes contaminated with Chemical, Biological, Radiological and Nuclear (CBRN) agents. The CBRN Crime Scene Modeler (C2SM) will use stereo cameras to create environment models and will be interfaced with a Directional Gamma Ray Probe, Chemical Agent Monitor and an Infra Red camera. The resulting 3D models will be augmented with readings from sensors indicating the threat level and distribution of contaminants in 3D. C2SM will be used either in a hand-held mode or in an automatic mode on-board of a mobile platform.

V. References

[1] Barfoot, T D, Se, S, Jasiobedzki, P: Visual Motion Estimation and Terrain Modelling for Planetary Rovers, Chapter in Intelligence for Space Robotics, Howard, A and Tunstel, E, Editors (Jet Propulsion Laboratory), TSI Press, Albuquerque, NM. 2006.

[2] David G. Lowe: Object recognition from local scale-invariant features, International Conference on Computer Vision, Corfu, Greece (September 1999), pp. 1150-1157

[3] David G. Lowe: Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[4] Point Grey Research www.ptgrey.com

[5] Stephen Se and Piotr Jasiobedzki: Instant Scene Modeler for Crime Scene Reconstruction. Proc. of IEEE Workshop on Advanced 3D Imaging for Safety and Security (A3DISS), San Diego, USA, June 2005.

[6] Stephen Se and Piotr Jasiobedzki: Photo-realistic 3D Model Reconstruction. Proc. of IEEE International Conference on Robotics and Automation, ICRA 2006, pp. 3076-3082, Orlando, Florida, May 2006.