INTERNATIONAL JOURNAL OF INTELLIGENT CONTROL AND SYSTEMS
VOL. 13, NO. 1, MARCH 2008, 46-57

Stereo-Vision Based 3D Modeling and Localization for Unmanned Vehicles

Stephen Se and Piotr Jasiobedzki

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

Instant Scene Modeler (iSM) is a vision system for generating calibrated photo-realistic 3D models of unknown environments quickly using stereo image sequences.

Equipped with iSM, unmanned military vehicles can capture stereo images and create 3D models to be sent back to the base station, while they explore unknown environments. Rapid access to 3D models will increase the operator situational awareness and allow better mission planning and execution, as the models can be visualized from different views and used for relative measurements.

Moreover, iSM also recovers the camera motion, also known as the visual odometry. As wheel odometry error grows over time, this can help improve the wheel odometry for better localization.

Apart from unmanned vehicles, iSM has also been used in forensics and mining applications, for creating 3D models of crime scenes and of the mine tunnels respectively.

***Index Terms*—Stereo computer vision, 3D modeling, Localization, Unmanned vehicles, Military, Security, Mining, Forensics**

1. INTRODUCTION

The creation of photo-realistic three-dimensional (3D) calibrated models of observed scenes has been an active research topic for many years. Such 3D models are very useful for both visualization and measurements in various applications such as planetary rovers, military, mining, forensics, archaeology, virtual reality, etc. The capability of creating 3D models automatically and quickly is particularly beneficial. A hand-held device is desirable in many situations as it can be used for scanning by simply moving it freely without any constraint on the motion. It can also be mounted on unmanned vehicles to create 3D models while it is exploring the environment.

We have developed the instant Scene Modeler (iSM) that is capable of quickly generating calibrated photo-realistic colour 3D models of unknown environments from a mobile stereo camera [1] . The system works in a hand-held mode where it can process image sequences and automatically stitch them together in 3D with no prior knowledge of the environment. The resulting 3D models can be visualized from different views and metric measurements can be performed on the models.

We have implemented a proof-of-concept iSM payload on an autonomous Unmanned Ground Vehicle (UGV) to provide photo-realistic 3D modeling and measurements. While the UGV autonomously explores the environment, the iSM payload will generate a photo-realistic 3D model of the environment which can be sent back wirelessly to the base station for mission reconnaissance.

From a remote location, the operator will have near real-time access to the 3D model of the unknown environment. The metrically accurate model may be augmented with multiple forms of additional sensory information, potentially including IR, or thermal imagery. Furthermore, as operationally needed, the iSM 3D modeling payload can be removed from the UGV and used in a hand-held mode by the operator.

Moreover, iSM can be used in forensics to create 3D models of the crime scenes, and it can be used in mining to create 3D models of the mine tunnels.

This paper is extended from "Stereo-Vision Based 3D Modeling for Unmanned Ground Vehicles" published at *SPIE Vol.6561: Unmanned Systems Technology IX*, Orlando, Florida, USA, Apr., 2007.

Stephen Se is with MDA, Space Missions, 9445 Airport Road, Brampton, ON L6S 4J3, Canada, phone 905-790-2800 x4270, e-mail: stephen.se@mdacorporation.com

Piotr Jasiobedzki is with MDA, Space Missions, 9445 Airport Road, Brampton, ON L6S 4J3, Canada.

2. PREVIOUS WORK

3D modeling has been a topic of intensive research for the last few decades. This section presents a brief overview of the main technologies: 3D acquisition, view registration. model construction, and a few 3D modeling systems applicable to unmanned vehicles.

2.1 3D Acquisition

The main approaches for depth acquisition include structured light, laser scanning and stereo. The structured light approach uses a projector to illuminate the object with patterns and recovers the 3D shape from a monocular image. It is effective for scanning objects but do not work well for scanning environments due to their limited range.

Blais [2] has recently reviewed the development of 3D laser imaging for the past 20 years. Auto-synchronous laser scanners can be used for both objects and environments due to their long depth of field and high accuracy at close range. Time-of-flight scanning laser rangefinders measure the time it takes for the light to travel to the object and back. Laser range scanners have to remain stationary during data acquisition and they are large, heavy, and tend to be expensive.

Stereo imaging is a passive technique and can recover the structure of the environment by matching features detected in multiple images of the same scene. It is very computationally intensive as the 3D data is computed from the images. The depth data could be noiser than the other approaches, as it relies on the natural texture on the surface and ambient lighting. Unlike laser scanners, cameras can capture complete images in microseconds, hence they can be used as mobile sensors or operate in dynamic environments. The cost, size, mass and power requirements of stereo cameras are much lower than those of scanning rangefinders.

2.2 View Registration

When multiple scans are obtained, they need to be registered together to build the 3D model. Registration can be carried out with a separate device that tracks the sensor or object position, or by matching the data sets manually or automatically.

The most common algorithm for automatic 3D data registration is Iterative Closest Point (ICP) algorithm [3] , which iteratively minimizes the distances between the overlapping regions of two sets of 3D points or surfaces. For vision systems, fiducials can be placed in the scene and the camera pose can be estimated by tracking these markers [4] . However, this involves changes to the environment and it is not possible for some applications. The capability to track natural features in the scene to recover camera motion is much preferred.

2.3 Model Construction

Registered 3D data sets contain redundant overlapping measurements and measurement noise. They contain often too much detail for efficient visualization and manipulation, and they need to be converted to other formats. One approach involves constructing geometrical models, e.g., 3D surfaces or volumes. Triangular meshes that consist of a large number of triangles are often used as they can represent complex surfaces.

The models can be obtained by creating surface meshes from individual views first and then stitching them together [5] . If there is a significant overlap between the individual views, this approach is rather inefficient due to the need for repeated stitching. The volumetric approach is more efficient as the 3D points are accumulated into voxel grid structures first. Only one triangular mesh is created for all the measurements using an iso-surface extraction algorithm, such as the marching cubes [6] . After the triangular mesh is generated, texture images are mapped to provide the photo-realism [7] .

2.4 3D Modeling Systems

3D modeling systems have been developed for city scanning or the large-scale reconstruction of urban scenes. Many of them use laser sensors, which can provide accurate 3D measurements directly at long ranges. We will focus on passive camera systems whose advantages have been described above. However, some of those systems require manual operation [8] [9] and hence, it is labour-intensive to create the models. Some other automatic 3D modeling systems simplify the scene as geometric primitives such as planes and polyhedra [10] or use generative building models [11] . Therefore, their application is limited to man-made environments, buildings and city blocks.

Military UGV application requires automatic 3D reconstruction and that the system works in all types of environments including outdoor natural terrains and caves. Pollefeys et al. [12] and Nister [13] presented systems which create 3D surface models from a sequence of images taken with a hand-held video camera. The camera motion is recovered by matching corner features in the image sequence. Dense stereo matching is carried out between the successive frames. The input images are used as surface texture to produce photo-realistic 3D models. Unlike stereo approaches, monocular approaches only output a scaled version of the original object. Moreover, it requires a long processing time.

Batch processing approaches such as bundle adjustment [14] may produce better 3D reconstruction as information from all the camera frames are optimized simultaneously. However, they are not suitable for unmanned vehicles which require real-time camera ego-motion estimation.

The objective of the ongoing DARPA Urbanscape project is to develop a real-time data collection and processing system for automatic geo-registered 3D reconstruction of urban scenes from video data [15] . Promising results were shown but it is far from being real-time. Multiple video streams as well as GPS and INS measurements are collected to reconstruct photo-realistic 3D models and place them in geo-registered coordinates.

3. INSTANT SCENE MODELER

iSM automatically creates 3D models from a mobile hand-held stereo camera. It computes the 3D data, estimates the camera motion and registers successive frames together. The user points the camera at a scene of interest and the system automatically creates a photo-realistic 3D calibrated model within minutes. Part of the processing includes computation of camera motion, which can be used for vehicle localization.

Figure 1 shows the architecture of the iSM system. Images are captured by the stereo camera and dense stereo disparity is computed for each stereo pair to obtain 3D data using known camera calibration. The system does not require any external sensors for computing the camera motion as it automatically extracts and tracks natural tie points in the images. The recovered camera motion is used to integrate 3D data obtained from the sequences. The 3D data is then converted to surface meshes, which are augmented by mapping texture from the colour images. The algorithms are described further in the next sections.

3.1 Tie Point Extraction

The system does not require any external sensors for computing the camera motion as it automatically extracts and tracks natural tie points in the images. We use a high level set of natural visual features called Scale Invariant Feature Transform (SIFT) as the tie points to compute the camera motion. SIFT was developed by Lowe [17] for image feature generation in object recognition applications.

The SIFT features are invariant to image translation, scaling, rotation, and partially invariant to illumination changes and to affine or 3D projections. These characteristics make them suitable as landmarks for robust matching when the cameras are moving around in an environment. The SIFT features are highly distinctive as each feature contains a 128-element local image descriptor.

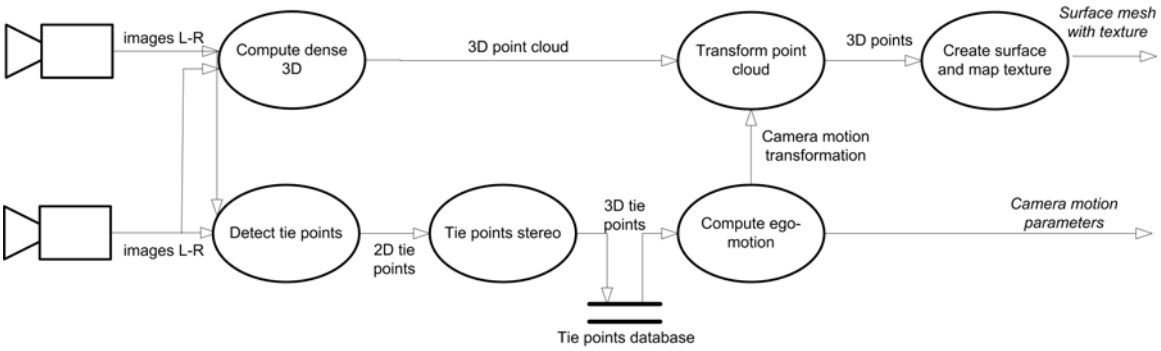

Figure 1 iSM system architecture

	Harris Corners	SIFT Features
Algorithm complexity	Easy to detect	Complex detection algorithm
Localization accuracy	Sub-pixel	Sub-pixel
Scales	Single or multiple scales	Multi-scale representation
Description	Image windows	Specific local image feature vector
Correspondence	Hard, many mismatches	Easy, few mismatches

Table 1 Comparison between Harris corners and SIFT features

Previous approaches to feature detection, such as the widely used Harris corner detector [18] , are sensitive to the scale of an image and therefore are less suitable for building feature databases that can be matched from a range of camera positions. A comparison between Harris corners and SIFT features is shown in Table 1.

3.2 Tie Points Stereo

Using the known stereo camera geometry, the SIFT features in the left and right images are matched using the following criteria: epipolar constraint, disparity constraint, orientation constraint, scale constraint, local feature vector constraint and unique match constraint [19] . The subpixel disparity d for each matched feature and hence the 3D coordinates (X, Y, Z) are also computed:

$$X = \frac{(c - c_0)I}{d}; Y = \frac{(r - r_0)I}{d}; Z = \frac{fI}{d}$$

where (r_0,c_0) are the image centre coordinates, I is the interocular distance and f is the focal length.

The stereo matched SIFT features in a lab scene are shown in Figure 2 (left), where the length of the line is proportional to the disparity. We can see that all the matches are consistent and correct. Typically, we obtain hundreds of SIFT 3D landmarks.

3.3 Camera Ego-motion Estimation

We recover the 6 degrees of freedom (dof) camera ego-motion when the camera moves freely in the hand-held mode. We employ a Simultaneous Localization And Mapping (SLAM) approach that uses the SIFT 3D landmarks to localize and simultaneously build a database map [19] . Instead of frame to frame matching, we match the SIFT features at each frame with the database to reduce error accumulation. Olson et al. [20] reported a 27.7% reduction in rover navigation error when multi-frame tracking is used, rather than considering each pair of frames separately.

Figure 2 Stereo matched SIFT features (left) and SIFT features that are matched to the database (right)

As the SIFT features are highly distinctive, they can be matched with very few false matches. This allows finding the camera movement that brings each projected SIFT landmark into the best alignment with its matching observed feature. To minimize the errors between the projected image coordinates and the observed image coordinates, we employ a weighted least-squares procedure to compute this 6 dof camera ego-motion.

Rather than solving directly for the 6 dof vector of camera ego-motion, Newton's method computes a vector of corrections $\boldsymbol{x}$ to be subtracted from the current estimate. No prediction model is used for the hand-held mode, and the previous camera pose is used as the initial estimate for the current frame.

Given a vector of error measurements **e** between the expected projection of the SIFT landmarks and the matched image positions, we would like to solve for an ***x*** that would eliminate this error:

$$WJx = We$$

where ***J*** is the Jacobian matrix $J_{i,j} = \partial e_i / \partial x_j$ and ***W*** is a diagonal matrix consisting of the inverse of the standard deviation of the measurements, assuming that landmarks are independent. As there are often more measurements than parameter, a weighted least-squares minimization is carried out taking into account the feature uncertainty.

The good feature matching quality implies a very high percentage of inliers, therefore outliers are simply eliminated by discarding features with significant least-squares errors. The minimization is repeated with the remainder matches to obtain the new correction terms.

Figure 2 (right) shows the SIFT features that are matched to the database. The line connecting the previous position to the current position is analogous to optical flow and we can see that all matches are consistent and correct.

3.4 Dense 3D Computation

Dense stereo disparity is computed from the left and right images using a correlation based algorithm to obtain 3D data. As with other stereo algorithms, the quality (accuracy, coverage and number of outliers) of the depth data depends on the presence of texture in the images.

3D data is computed in the camera reference frame and is transformed using the camera ego-motion estimated for this frame. Typically, the initial camera pose is used as the reference.

3.5 Mesh Creation

Using all 3D points obtained from the stereo processing is not efficient as there are a lot of redundant measurements, and the data may contain noise and missing regions. Representing 3D data as a triangular mesh reduces the amount of data when multiple sets of 3D points are combined. Furthermore, creating surface meshes fills up small holes and eliminates outliers, resulting in smoother and more realistic reconstructions.

To generate triangular meshes as 3D models, we employ a voxel-based method [6] , which accumulates 3D points into voxels at each frame with their associated normals. It creates a mesh using all the 3D points, fills up holes and works well for data with significant overlap. It takes a few seconds to construct the triangular mesh at the end, which is dependent on the data size and the voxel resolution.

3.6 Texture Mapping

The photo-realistic appearance of the reconstructed scene is created by mapping camera images as texture. Such surfaces are more visually appealing and easier to interpret as they provide additional surface details. Colour images from the stereo camera are used for texture mapping. As each triangle may be observed in multiple images, the algorithm needs to select a texture image for each triangle. To reduce the appearance of seam lines between triangles, we select the texture images that can cover the most number of triangles. The model will therefore need the minimal number of texture images, and hence this allows faster model loading and lower storage requirement.

4. IMPLEMENTATIONS

We have developed several iSM prototype systems:

- Hand-held
- Vehicle-mounted
- Motorized

4.1 Hand-held system

The hand-held version of iSM is shown in Figure 3. The main hardware components of iSM are a stereo camera and a computer. We currently use a colour Bumblebee stereo camera from Point Grey Research (PGR) [16] . It is a firewire camera that can capture up to 15 frames per second. iSM 3D processing software can run on any PC equipped with a firewire interface.

Figure 3 Hand held iSM

In one of the experiments, we modeled a façade of a house. The camera was moved freely pointing at different portions of the house and about 30 seconds of 640x480 resolution stereo images were captured. Figure 4 shows two of the images from a sequence. iSM processed these images automatically and created a photo-realistic 3D model in around 5 minutes on a Pentium IV 2.4GHz laptop.

Figure 4 Two images from the house sequence

The output 3D model is stored in the VRML (Virtual Reality Modeling Language) format. The user can navigate in the 3D model, and view it from any direction and distance. Figure 5 shows two views of the 3D model. We can see that iSM can reconstruct the overall 3D model by integrating all the input images, each of which captured with a limited field of view.

More advanced visualization and user interaction is provided in our visualization GUI (Graphical User Interface). As the 3D model is calibrated, the user can perform measurements (such as distance, angle, area) on the 3D model and the user can also annotate the model, as shown in Figure 6 (left). The camera trajectory recovered from the ego-motion estimation can be visualized, as shown in Figure 6 (right). The red, green, blue axes correspond to the X, Y, Z axes of the camera respectively. Moreover, the GUI also provides other features such as movie creation using trajectories defined with keyframes, model alignment and the export of 3D models into DXF format.

Figure 5 Two views of the 3D model created by iSM for the house scene

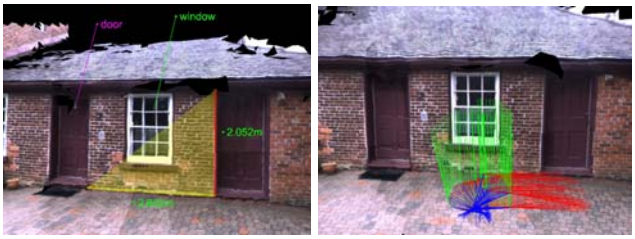

Figure 6 iSM visualization GUI showing the annotation and measurement (left) and the recovered camera trajectory (right)

4.2 Vehicle-mounted system

For autonomous vehicles and planetary rovers, the creation of 3D terrain models of the environment is useful for visualization and path planning [21] . Apart from the 3D model, iSM also computes the camera motion (this estimation is also known as visual odometry) which allows the vehicles to localize themselves more accurately, as wheel odometry is prone to errors due to wheel slippage [22] . iSM can be deployed on both tele-operated and autonomous vehicles to create 3D models while the vehicles traverse in unknown environments.

One of the MDA autonomous test vehicles is shown in Figure 7. The chassis of this rover is an iRobot ATRVJr with a custom vision system [21] . The stereo camera was constructed using a pair of Sony DFW-X700 cameras, mounted on a rigid bar and affixed to a pan-tilt unit. The camera field of view is approximately 45 degrees horizontal and 35 degrees vertical. There are currently two computers on board, a dual Pentium III 1 GHz with 1 GB of RAM (inside the red box) and a dedicated vision computer consisting of a Pentium M 1.8 GHz with 1 GB of RAM. The vision computer also houses our hardware accelerated vision processing boards, a Tyzx DeepSea2 [23] for dense stereo calculations and an AlphaData ADM-XRC board with a Virtex II Xilinx FPGA (Field Programmable Gate Array) running our implementation of SIFT feature extraction [21] . There are various other sensors onboard as well: sonar rangefinders, SICK laser rangefinder, DGPS, compass, inertial measurement unit and inclinometer.

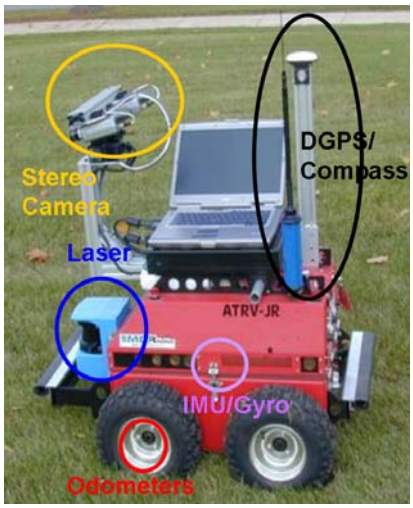

Figure 7 Vehicle mounted iSM

iSM has been tested on this testbed traversing at a desert in Nevada. Figure 8 (left) shows an image from a sequence captured by the vehicle during traversal of 40m and Figure 8 (right) shows the reconstructed 3D terrain model with a virtual vehicle inserted for visualization. Figure 9 shows the recovered camera trajectory without using wheel odometry. Despite the jerky motion, iSM is able to compute the camera motion and create a photo-realistic 3D model from the input stereo image sequence.

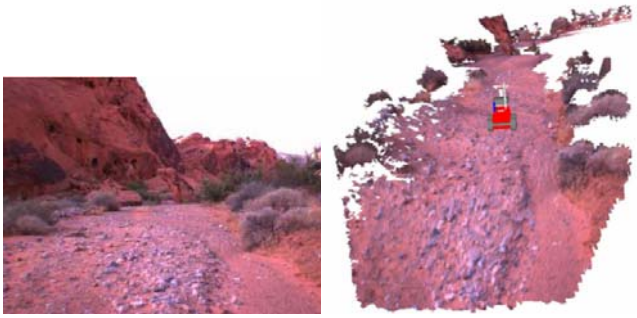

Figure 8 An image from a sequence taken by the MDA Autonomous Vehicle at a desert in Nevada (left) and the reconstructed 3D model with the vehicle model inserted for visualization (right).

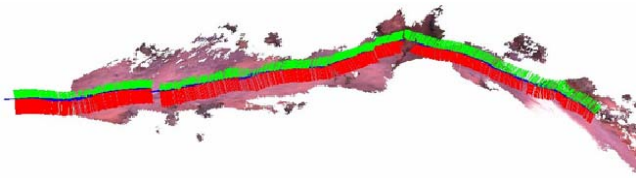

Figure 9 Camera trajectory computed by iSM for the Nevada dataset

The on-board FPGA computes SIFT features at rates of 7 Hz for 1024x768 resolution images. This allows 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 120m 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. When operating on vehicles, the ego-motion estimates are fused with the measurements from wheel odometry and inertial sensors.

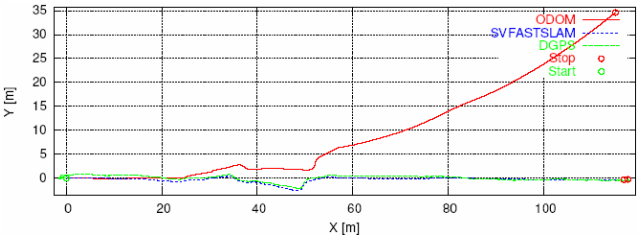

Figure 10 Comparison of errors incurred by wheel and visual odometry

4.3 Motorized system

The motorized tripod-mounted version of the system is used when automatic image acquisition and full scene coverage are required. A commercial system for modeling underground mines is shown in Figure 11. A stereo camera with a longer baseline allows imaging of objects at longer distances and an integrated camera light provides the necessary illumination. The camera head is mounted on a motorized pan-tilt-unit and moves to pre-programmed positions, recording images and telemetry.

Figure 11 instant Mine Modeler – a motorized version of iSM

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 in Figure 12. Multiple scans have been acquired from different positions and registered together using the overlapping sections. Additional information on the commercial system for mining is provided in Section 5.3.

Figure 12 Reconstruction of an underground tunnel (tunnel view and top-down view)

5. Modeling Enhancements

5.1 Auto-Referencing

When a continuous scan is not possible but separate scans are collected, we will create multiple 3D models. As each sequence starts at an unknown location and hence has a different origin, we need to align the multiple meshes to merge them together. We refer this process of aligning multiple meshes automatically as auto-referencing [1] . This capability is useful for creating 3D model for larger environments, as it can automatically combine individual 3D models that cover smaller environments.

We use highly distinctive SIFT features during automatic alignment, as each 3D model has an associated SIFT database map. The problem is defined as, given two SIFT database maps without prior information about their relative position, estimate their 6 dof alignment, provided that there are sufficient overlapping features between them.

Map alignment using SIFT features has been proposed in [24] , but it is limited to 3 dof as the mobile robot can only undergo more or less planar motion. For iSM, the user can start the camera at any position and hence this is extended to 6 dof. Instead of the Hough Transform approach, we employ a RANSAC approach which is more efficient, especially with the higher dimensions.

The algorithm involves the following steps:

• Create a list of tentative matches. For each SIFT feature in the second database, find the feature in the first database which matches best, in terms of the SIFT local image vector

• Randomly select 3 tentative matches from the list and compute the 6 dof alignment parameters from them

• Seek support by checking all the tentative matches that support this particular alignment

• Repeat this random selection, alignment computation and support seeking process many times. The alignment with most support is our hypothesis.

• Proceed with a least-squares minimization for the inliers which support this hypothesis and obtain a better estimate for the alignment

The probability of a good sample τ for RANSAC [25] is given by:

$$\tau = 1 - (1 - (1 - \varepsilon)^p)^m$$

where ε is the contamination ratio (ratio of false matches to total matches), p is the sample size and m is the number of samples required. In this case, with p = 3, ε = 0.8, τ = 99%, m = 573. That is, for 99% probability of a good sample, we need to sample at least 573 times. The algorithm works well if there is sufficient number of overlapping SIFT features.

Once we have found the alignment based on the SIFT database maps, we can put the two 3D models together to obtain a more complete reconstruction of the environment.

An example with two separate image sequences captured in the lab and two 3D models with overlapping regions is shown in Figure 13. By applying the auto-referencing algorithm, the 6 dof alignment has been recovered based on the two SIFT database maps and the two models can be put together automatically.

Figure 13 (top) shows the two individual 3D models. Figure 13 (bottom) shows the aligned model after auto-referencing which took less than 3 seconds on a Pentium IV 2.8 GHz processor. It can be seen that the correct alignment is obtained without any user interaction. ICP is commonly used for aligning multiple sets of 3D laser scans, but it requires an initial estimate of the relative alignment, or user interaction is needed to establish several correspondences. The developed SIFT-based auto-referencing algorithm aligns two 3D models automatically; multiple 3D models can be aligned in an incremental or pair-wise fashion.

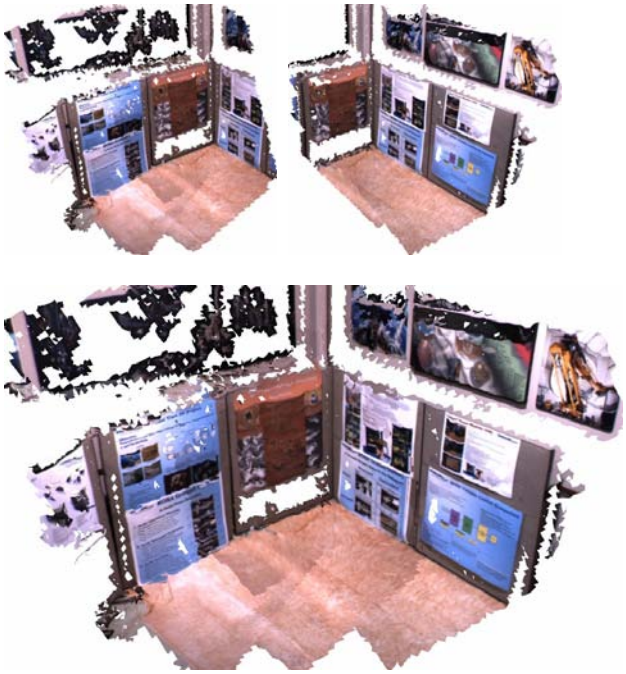

Figure 13 Two 3D models with some overlap but unknown alignment. (top) and the aligned model after auto-referencing (bottom)

5.2 Pattern Projector

Indoor man-made environments may not have much texture for high coverage dense stereo matching. In order to obtain better 3D models, we have built a flash pattern projector that projects a random dot pattern onto the scene, as an optional add-on to the stereo camera [1] . The projector is synchronized with the stereo camera and the projector trigger is controlled via software.

SIFT features found in the flash image should not be used to compute the camera motion, as the same pattern is projected all the time. Therefore, we interleave normal images with a flash image every 10 frames. The ego-motion estimation for the normal images is the same as before, but the camera location for the flash images are interpolated from the SIFT-based ego-motion of the normal images.

Figure 14 shows an example of a normal image and a flash image showing the random dot pattern of an indoor scene. Figure 15 shows the screenshots of two 3D models, one without the flash and one with the flash. The black regions indicate areas where there is not enough texture to recover the depth. It can be seen that the coverage of the 3D model is increased substantially with the pattern projector.

Figure 14 Sample image without flash (left) and with flash (right)

Figure 15 Screenshots of 3D model without (left) and with the pattern projector (right)

6. APPLICATIONS

5.1 Military

Special Operations Forces (SOF) are often required to enter unknown and potentially hostile environments with no prior knowledge of the layout or possible deployment of hostile forces. Lacking situational awareness prior to entry, the SOF can be taken by surprise by hostile forces, explosives, or other threats. The survivability and effectiveness of the SOF will be greatly enhanced by remotely acquiring knowledge of the environment's layout. The use of UGVs keeps the SOF personnel out of danger and allows them to carry out other critical activities while

the UGV is autonomously modeling the environment. With the acquired knowledge of the environment, the SOF will be able to carry out more strategic, targeted and safe operations.

iRobot PackBot [26] is a highly-robust, all-weather, all-terrain, man-portable UGV platform, equipped with two main treads for locomotion and two articulated flippers with treads to climb over obstacles. PackBot can travel at sustained speeds of up to 4.5 mph. It is 27 inches long, 16 inches wide, and 7 inches tall, and weighs 40 pounds. All electronics including the on-board computer are inside a compact, hardened enclosure. Each PackBot can withstand a 400G impact, the equivalent of being dropped from a second storey window onto concrete. Each PackBot is also waterproof to 3 metres.

PackBot is at home in both wilderness and urban environments. In the wilderness, PackBot can drive through fields and woods, over rocks, sand, and gravel, and through water and mud. In the city, PackBot can drive on asphalt and concrete, climb over curbs, and climb up and down stairs while carrying a payload.

While the PackBot is tele-operated, autonomous urban navigation capabilities have been developed in the Wayfarer project [27] . A modular navigation payload has been developed that incorporates a 3D stereo vision system, a 360-degree planar LIDAR, GPS, INS, compass, and odometry. This payload can be attached to any PackBot to provide the robot with the capability to perform autonomous urban reconnaissance missions. The PackBot with Wayfarer technology will be able to scout unknown territory and send back occupancy maps along with video image sequences.

The Wayfarer navigation payload includes software components for obstacle avoidance, building perimeter and urban street following, and map-building. The obstacle avoidance system enables the PackBot to avoid collisions with a wide range of obstacles in both outdoor and indoor environments. This system combines 360-degree planar LIDAR range data with 3D obstacle detection using stereo vision. A real-time Hough transform is used to detect linear features in the range data that correspond to building walls and street orientations. The LIDAR range data builds an occupancy grid map of the robot's surroundings in real-time. Data is transmitted via UDP over wireless Ethernet to an OpenGL-based Operator Control Unit (OCU) that displays this information graphically and in real-time.

We have mounted iSM as a proof-of-concept payload on the Wayfarer PackBot, as shown in Figure 16. The payload consisted of a stereo camera and a compact laptop computer (Toshiba Libretto with 1.1 GHz processor). The stereo camera with the MDA logo belonged to the iSM payload while the second one was part of Wayfarer's sensor package. The iSM payload camera was pointed slightly to the right side during the test runs, so that it could get a better view of the building for 3D reconstruction, rather than capturing the road ahead only.

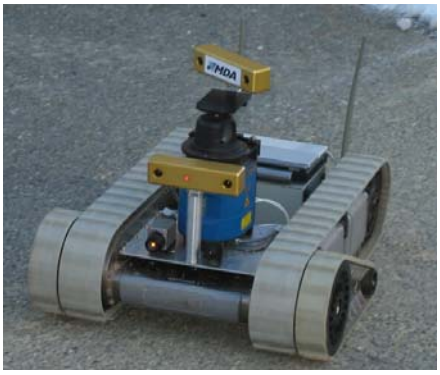

Figure 16 A proof-of-concept iSM 3D modeling payload mounted on the iRobot Wayfarer PackBot. The payload consists of a stereo camera (with the MDA logo) and a compact laptop at the back.

There was no data interface between the iSM payload computer and the PackBot computer. The stereo images were captured and processed by the payload computer on the PackBot. A separate control computer was used for communication with the iSM payload and 3D model visualization.

Due to the limited payload computing resources and the relatively high speed of Wayfarer PackBot, the images were captured with a lower resolution of 320x240 pixels and at 7Hz. During a test run, the Wayfarer PackBot autonomously navigated around the building in the perimeter following mode while iSM was capturing images. At the end of the test run, iSM started processing the stereo images and created a photo-realistic 3D model.

Figure 17 shows selected images captured by the iSM payload for test run 1. The traverse was around 20m and took around 1 minute. The processing time was under 5 minutes on the payload computer. Screenshots of the resulting 3D model from different views are shown in Figure 18, together with the recovered camera trajectory.

Figure 17 Selected images from test run 1

Another test was carried out in which the Wayfarer PackBot turns around the corner of the building. Figure 19 shows some input images captured by the iSM payload for test run 2. Screenshots of the resulting 3D model from different views are shown in Figure 20. The total traverse is around 40m and takes around 3 minutes. The processing time is less than 10 minutes.

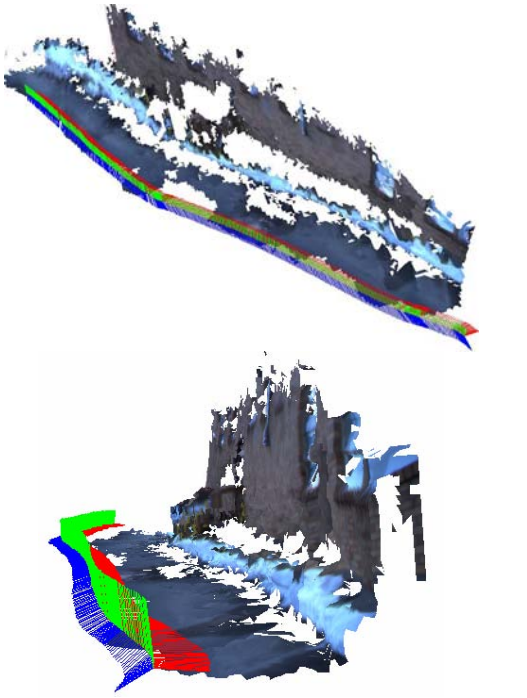

Figure 18 Different views of the resulting 3D models with recovered trajectory highlighted for test run 1

Figure 19 Selected images from test run 2

Key capabilities of a mobile robot system are simultaneous localization and building maps of visited environments. Experimental results show that the iSM payload can complement the Wayfarer PackBot in both aspects. While the Wayfarer PackBot is autonomously following the building perimeters and avoiding obstacles, it builds 2D occupancy grid maps from the laser sensor, whereas iSM is capable of creating photo-realistic 3D models. The photo-realistic 3D models provide better situational awareness than 2D occupancy grid maps and can be used for change detection.

The Wayfarer PackBot uses wheel odometry for localization and to build the occupancy grid map. As wheel odometry is prone to error, an additional INS/GPS unit is used to improve localization [28] . iSM recovers the camera motion, also known as the visual odometry, which can be fused with the wheel odometry for better localization. However, as the iSM processing was not done in real-time (real-time visual odometry requires the SIFT-FPGA processor described in Section 4.2) and there was no communication between iSM and PackBot computers, the PackBot could not use the visual odometry produced by iSM.

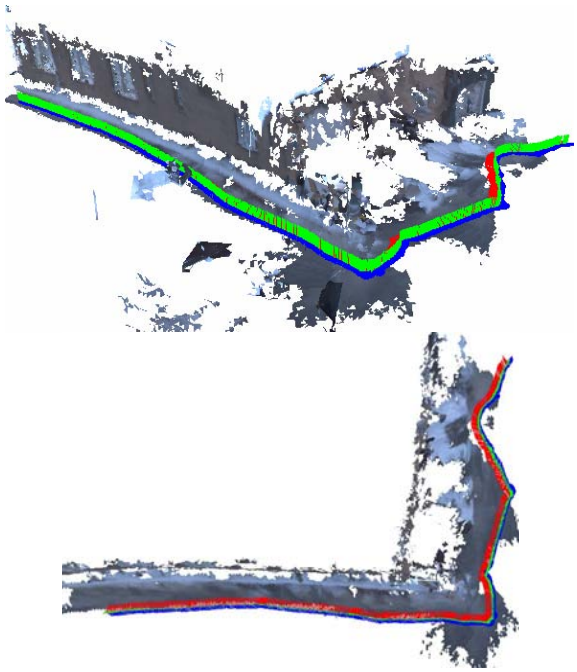

Figure 20 Different views of the resulting 3D models with recovered trajectory highlighted for test run 2

As the effectiveness of the SOF will be greatly enhanced by having a layout of the environment, current military operations of UGVs in urban warfare threats involve the operator hand-sketching the environment from live video feed. iSM eliminates the need for an additional operator as the 3D model is generated automatically. The safety of the SOF will be increased through the remote operation and monitoring from a secure stand-off location. As a result, overall mission effectiveness, success, and safety will be greatly increased.

5.2 Security and Forensics

Documenting crime scenes is a tedious process that requires the investigators to record vast amounts of data using video and still cameras, measuring devices, taking samples and recording observation. All this data must be later stored in structured manner for easy access during investigations.

With iSM, investigators can create a 3D model of the crime scene quickly without much disturbance to the crime scene [29] . Unlike traditional 2D imaging, measurements can be performed on the 3D model. As our 3D model is

fully calibrated, there is no need to measure a reference object in the scene to scale the model. The police can also perform additional measurements they may have missed using the 3D model after the crime scene is released.

The 3D model can be shown to other officers who have not been to the crime scene. Apart from helping the police investigation, the 3D model can potentially be shown in court so that the judge and the jury can understand the crime scene better.

Figure 21 Two images from the mock crime scene sequence.

Figure 21 shows selected images obtained with the handheld camera at a mock crime scene set up in our lab, while the operator was moving the camera around in the scene. The 640x480 image sequence was captured approximately within 1 minute and the processing took around 10 minutes. Figure 22 shows the 3D model generated from the input sequence.

Figure 22 Reconstructed 3D model of the mock crime scene.

Investigating crime scenes where Chemical, Biological, Radiological and Nuclear (CBRN) agents have been deployed poses great dangers to first responders. Any prior decontamination of a crime scene may result in destruction of potentially vital evidence. Technologies that reduce the need to enter the scene or to reduce exposure of first responders are essential.

In an on-going project, we are developing a CBRN Crime Scene Modeler (C2SM), a 3D modeling system for CBRN contaminated scenes, based on iSM. C2SM uses stereo cameras to create 3D models and interfaces with a Directional Gamma Ray Probe, Chemical Agent Monitor and an Infra Red camera [30] . The resulting 3D models will be augmented with readings from sensors indicating the threat level and distribution of contaminants in 3D.

C2SM operates either as a hand-held device or in a robot mode. In the hand-held mode, the operator uses the system similarly to a video camera to acquire the images and data. In the robot mode, C2SM operates on board of a mobile platform as shown in Figure 23 and is controlled remotely from an operator station. The data is processed in an embedded computer and models are available within minutes. The multi-modal models are visualized in 3D and may be augmented with annotations and additional information. All this information is stored in an event database and transferred to a command centre.

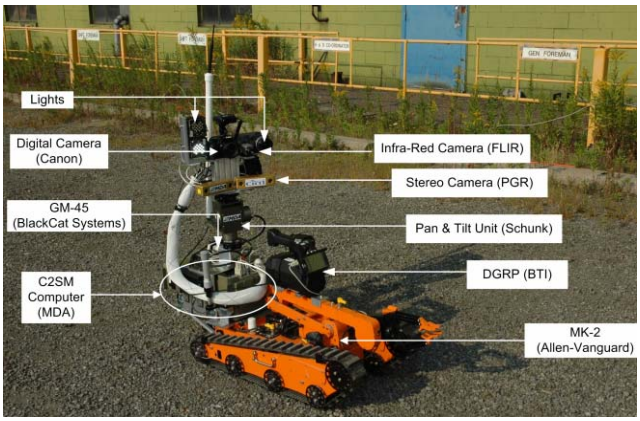

Figure 23 C2SM prototype for CBRN crime scene investigation mounted on a mobile robot

5.3 Mining

Photo-realistic 3D models are useful for survey and geology in underground mining. Currently, mining companies can only survey the mine from time to time to keep track of the mine advance due to the cost of surveys. With our 3D models, the mine map can be updated after each daily drill/blast/ore removal cycle to minimize any deviation from the plan. In addition, the 3D models can also allow the mining companies to monitor how much ore is taken at each blast.

Geological mapping are typically not carried out by geologists at every drill cycle as the cost would be prohibitive. This does not allow monitoring of the ore content or adapting the mine exploration plan to local conditions. As our system can be operated by a crew that is already on site, the images can be collected daily and 3D models may be sent to geologists on the surface.

We have developed a version of the system specifically for underground mining – instant Mine Modeler (iMM) shown in Figure 11. iMM was tested in several underground mines. Mine surfaces provide rich features for both feature extraction and dense stereo matching. Figure 24 shows a 3D model reconstructed at an underground mine cavity using the motorized mode. Geological annotation has been overlaid on the 3D model and these 3D geological features can be exported into mine management software to model the ore body better.

All the 3D models need to be transformed into the mine coordinates, in order to be placed in the right location of the overall mine map. Accurate registration with the mine coordinate system is performed by measuring the camera location relative to survey markers for each scan. A laser rangefinder integrated in the iMM head provides highly accurate range measurements for distances up to 100m.

Past and recent as-built models can be linked together and used for planning the exploration and monitoring the tunnel advancement. Figure 25 shows consecutive scans of an advancing mine face. Volume between consecutive faces can be calculated to estimate how much ore has been taken at each round.

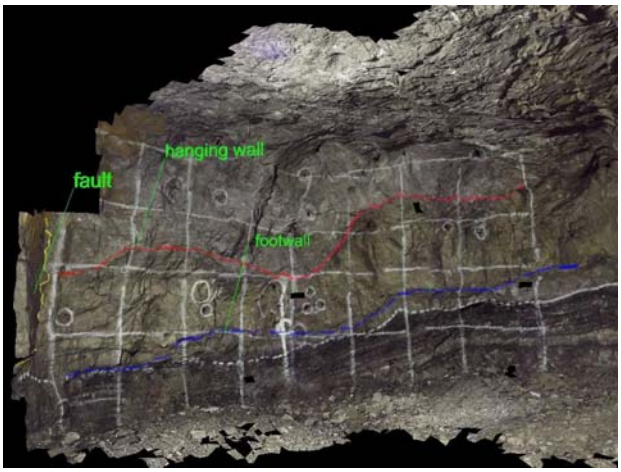

Figure 24 3D model of underground mine overlaid with geological annotation

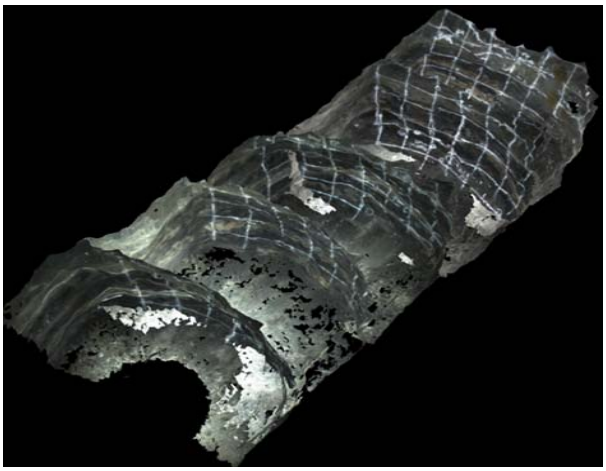

Figure 25 As-built mine model showing tunnel advancement

7. CONCLUSIONS

In this paper, we have presented a family of 3D modeling systems based on the instant Scene Modeler (iSM). The approach relies on the processing of image sequences 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 photo-realistic 3D models. When implemented on dedicated hardware (FPGA), the motion is estimated at several frames per second allowing it to be used for visual odometry and vehicle localization. It creates photo-realistic 3D calibrated models of environments automatically within minutes.

We have developed three versions of iSM for various applications. The motorized version is particularly suitable for mining as it allows automatic image acquisition, full scene coverage and registration within mine coordinate system. The models are annotated with 3D geological features that enable ore body mapping and provide daily as-built update of the mine map. A rugged commercial production system (iMM) has recently been developed for creating 3D models at underground mines. Future work includes generating 3D models while the mining vehicle is traversing in the mine.

The hand-held version offers flexibility for forensic investigators. Photo-realistic reconstruction of crime scenes combined with high resolution images and operator annotations helps collecting and storing detailed evidence collected during investigations. Current work focuses on adaptation of the 3D modeling technologies for investigating scenes contaminated with CBRN agents. The system deployed on a remotely-controlled mobile platform reduces exposure of first responders to dangerous agents. Incorporation of chemical and gamma radiation detectors allows mapping of contamination levels and provides additional situational awareness to the first responders.

The vehicle-mounted version can be used for military application on-board an UGV. The SOF situational awareness before entering an unknown environment will be greatly enhanced through the acquisition of a high-fidelity, photo-realistic 3D model. Future work includes better integration of the iSM payload with the PackBot. Interfacing the payload and the PackBot will allow iSM to make use of PackBot wheel odometry, improve it robustly with visual information, and send the new estimated .location to the robot controller. iSM processing can be optimized to provide online visual odometry by means of software and hardware acceleration.

Apart from ground vehicles, iSM can potentially be deployed on UAV (Unmanned Aerial Vehicles) and UUV (Unmanned Underwater Vehicles) for 3D modeling.

In general, using images alone for motion estimation causes accumulation of error over long sequences. It can be noticed that the building wall looks slightly curved in Figure 20. Backward correction techniques for map building are considered to improve the model correctness for long sequences [24] . Integration with global sensors such as Global Positioning System or radio beacons will allow resetting the error at selected locations where such data will be available.

ACKNOWLEDGEMENT

We thank Brian Yamauchi, Chris Jones and Erik Schoenfeld at iRobot Corporation for their help with the PackBot experiments and discussions. We acknowledge contributions of the MDA staff working on the iMM, C2SM and Autonomous Vehicles projects. We thank Bovaird House and Andrew Roth for the field tests.

References

[1] S. Se and P. Jasiobedzki, Photo-realistic 3D model reconstruction, IEEE International Conference on Robotics and Automation (ICRA), pages 3076-3082, Orlando, Florida, May 2006.

[2] F. Blais, Review of 20 years of range sensor development, Journal of Electronic Imaging, 13(1):231-240, Jan 2004.

[3] P. Besl and N. McKay, A method for registration of 3-D shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(2):239-256, 1992.

[4] U. Neumann and Y. Cho, A self-tracking augmented reality system, ACM International Symposium on Virtual Reality and Applications, pages 109-115, July 1996.

[5] G. Turk and M. Levoy, Zippered polygon meshes from range images, SIGGRAPH'94, pages 311-318, Orlando, Florida, July 1994.

[6] G. Roth and E. Wibowo, An efficient volumetric method for building closed triangular meshes from 3-D images and point data, Graphics Interface (GI), pages 173-180, Kelowna, B.C., Canada, 1997.

[7] M. Soucy, G. Godin, and M. Rioux, A texture-mapping approach for the compression of colored 3D triangulations, The Visual Computer, 12:503-514, 1996.

[8] P. Debevec, C.J. Taylor, and J. Malik, Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach, SIGGRAPH, pages 11-20, 1996.

[9] C. Rother and S. Carlsson, Linear multi-view reconstruction and camera recovery using a reference plane, International Journal of Computer Vision, vol. 49, no. 2-3, pages 117-141, 2002.

[10] T. Werner and A. Zisserman, New techniques for automated architectural reconstruction from photographs, European Conference on Computer Vision, volume II, pages 541-555, 2002.

[11] A.R. Dick, P.H.S. Torr, and R. Cipolla, Modelling and interpretation of architecture from several images, International Journal of Computer Vision, vol. 60, no. 2, pages 111-134, 2004.

[12] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, Visual modeling with a hand-held camera, International Journal of Computer Vision, vol. 59, no. 3, pages 207-232, 2004.

[13] D. Nister, Automatic passive recovery of 3D from images and video, International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), pages 438-445, 2004.

[14] A.W. Fitzgibbon and A. Zisserman, Automatic Camera Recovery for Closed or Open Image Sequences, European Conference on Computer Vision (ECCV), pages 311-326, Germany, 1998.

[15] A. Akbarzadeh, J.M. Frahm, P. Mordohai, B. Clipp, C. Engels, D. Gallup, P. Merrell, M. Phelps, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewenius, R. Yang, G. Welch, H. Towles, D. Nister, and M. Pollefeys, Towards urban 3D reconstruction from video, International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), 2006.

[16] Point Grey Research, http://www.ptgrey.com

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

[18] C.J. Harris and M. Stephens, A combined corner and edge detector, 4th Alvey Vision Conference, pages 147-151, Manchester, 1988.

[19] S. Se, D. Lowe, and J. Little, Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks, International Journal of Robotics Research, vol. 21, no. 8, pages 735–758, August 2002.

[20] C.F. Olson, L.H. Matthies, M. Schoppers, and M.W. Maimone, Robust stereo ego-motion for long distance navigation, IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pages 453-458, June 2000.

[21] T. Barfoot, S. Se and P. Jasiobedzki, Vision-based localization and terrain modeling for planetary rovers, Intelligence for Space Robotics, editors A. Howard and E. Tunstel, TSI Press, Albuquerque, NM, 2006.

[22] M. Maimone, Y. Cheng and L. Matthies, Two years of visual odometry on the Mars exploration rovers, Journal of Field Robotics, 24(3):169-186, 2007.

[23] Tyzx, http://www.tyzx.com

[24] S. Se, D. Lowe, and J. Little, Vision-based global localization and mapping for mobile robots, IEEE Transactions on Robotics, vol. 21, no. 3, pages 364-375, June 2005.

[25] M.A. Fischler and R.C. Bolles, Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography, Commun. ACM, 24:381-395, 1981.

[26] iRobot, http://www.irobot.com

[27] B. Yamauchi, The Wayfarer modular navigation payload for intelligence robot infrastructure, SPIE Vol. 5804: Unmanned Ground Vehicle Technology VII, Orlando FL, March 2005.

[28] B. Yamauchi, Autonomous urban reconnaissance using man-portable UGVs, SPIE Vol. 6230: Unmanned Systems Technology VIII, Orlando, FL, April 2006.

[29] S. Se and P. Jasiobedzki, Instant scene modeler for crime scene reconstruction, IEEE Workshop on Advanced 3D Imaging for Safety and Security (A3DISS), San Diego, USA, June 2005.

[30] P. Jasiobedzki, H. Ng, M. Bondy, C. McDiarmid, Three-dimensional modeling of environments contaminated with chemical, biological, radiological, and nuclear (CBRN) agents. SPIE Vol. 6943, Orlando, March 2008.

Dr Stephen Se is with the Research and Development department at MDA in Canada, developing computer vision systems for space and terrestrial applications. He received B.Eng. degree with first class honours in Computing at Imperial College of Science, Technology and Medicine, University of London in 1995 and a D.Phil. degree in the Robotics Research Group at the University of Oxford in 1999. He then worked as a post-doctoral researcher at the University of British Columbia (1999-2001) on vision-based mobile robot localization. His research interests include computer vision, mobile robotics, 3D modeling, image processing and artificial intelligence.

Dr Piotr Jasiobedzki is a staff scientist at MDA in Brampton, Canada, where he leads projects on advanced vision systems for space and terrestrial applications. His research expertise includes 3D computer vision, sensor fusion, and robot control and navigation. Piotr has two patents and published over 40 papers.