Photo-realistic 3D Model Reconstruction

Stephen Se and Piotr Jasiobedzki
MDA, Space Missions
9445 Airport Road, Brampton, Ontario, L6S 4J3, Canada
{stephen.se,piotr.jasiobedzki}@mdacorporation.com

Abstract

Photo-realistic 3D modeling is a challenging problem and has been a research topic for many years. Quick generation of photo-realistic three-dimensional calibrated models using a hand-held device is highly desirable for applications ranging from forensic investigation, mining, to mobile robotics. In this paper, we present the instant Scene Modeler (iSM), a 3D imaging system that automatically creates 3D models using an off-the-shelf hand-held stereo camera. The user points the camera at a scene of interest and the system will create a photo-realistic 3D calibrated model automatically within minutes. Field tests in various environments have been carried out with promising results.

1. Introduction

Creation of photo-realistic three-dimensional (3D) calibrated models of observed scenes and objects has been an active research topic for many years. Such 3D models can be used for both visualization and measurements, in many applications including planetary rover exploration, forensics, mining, geology, archaeology, virtual reality, etc.

Object scanning and environment modeling can be regarded as two types of 3D modeling, as outside-looking-in approach is more suitable for object scanning whereas inside-looking-out approach is more suitable for environment modeling. 3D data can be obtained using various rangefinders, computed from stereo images or monocular sequences. The raw 3D measurements are then converted to representations suitable for display and manipulation.

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. This allows the user to position the sensor relative to the scene to get the optimal view and full coverage. The capability of creating 3D models automatically and quickly is particularly beneficial.

2. Previous Work

3D modeling has been a topic of intensive research for the last few decades. Instead of providing a comprehensive literature survey, we will review some previous work on 3D modeling related to iSM in this section. The key components of iSM include depth acquisition, view registration and model construction. Moreover, we will look into several other hand-held systems and compare them with iSM.

2.1. Depth 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 very 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. Autosynchronous 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 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 separate device can be a tracker such as Polhemus FastScan [12] or a turntable on which the object is placed such as Cyberware Model Shop [3].

Photogrammetry techniques (e.g., PhotoModeler [11]) can be used to create 3D models from sequences of monocular images, by manually establishing correspondences between features in different images to estimate their 3D coordinates. However, it is a time consuming process, limited to simple objects with polygonal faces, and does not recover the scale of the model.

The most common algorithm for automatic 3D data registration is Iterative Closest Point (ICP) algorithm [1], which iteratively minimizes the distances between the overlapping regions of two set 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 [8]. 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 details 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 [22]. If there is a significant overlap between the individual views, this approach is rather inefficient due to the need for repeated stitching. Volumetric approach is more efficient in such situations as the 3D points are accumulated into voxel grid structures first. Then only one triangular mesh is created for all the measurements using an iso-surface extraction algorithm, such as the marching cubes [16]. After the triangular mesh is generated, texture images are mapped to provide the photo-realism [21].

2.4. Hand-held Devices

Pollefeys *et al.* [13] and Nister [9] 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. However, it requires a long processing time and outputs a scaled version of the original object.

Hebert [6] proposed a self-referenced sensor which consists of two cameras and a cross-hair laser light projector. Frame to frame registration is achieved using a set of fiducials projected with an additional stationary laser system. The system requires a long acquisition time as it can capture only sparse 3D data for each frame and the 3D models do not have photo-realistic appearance.

Rusinkiewicz *et al.* [17] presented a real-time 3D modeling system that allows the user to rotate an object by hand and see a continuously-updated model as the object is scanned. It consists of a 60Hz structured-light rangefinder and a real-time variant of ICP for alignment. It is limited to the outside-looking-in case and does not acquire colour.

Popescu *et al.* [14] proposed the ModelCamera, which is a low-cost hand-held scene modeling device. It consists of a digital video camera with 16 laser pointers attached to it. ModelCamera acquires 16 depth samples per frame and registers the frames using depth and colour information. The surfaces are approximated with a few quadratics and this approach only works for smooth continuous surfaces.

ISM uses stereo cameras to obtain 3D data, estimate camera motion and register successive frames together. The resulting models are fully calibrated (allow Euclidean measurement) and have photo-realistic appearance. The data acquisition and processing takes minutes.

3. Instant Scene Modeler

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) [15] at 640x480 image resolution. It is a firewire camera and can capture up to 15 frames per second. iSM software can run on any PC with firewire interface.

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 camera images.

3.1. Dense Stereo

The left and right images are matched to obtain dense disparity data, which is then converted to 3D depth data. We run PGR's optimized Triclops library for correlation-based dense stereo. 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.

3.2. SIFT Extraction

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 [7] for image feature generation in object recognition applications. Apart from the location, scale and orientation, each SIFT feature is also described by a local image vector for high specificity.

The features are invariant to image translation, scaling, rotation, and partially invariant to illumination changes and affine or 3D projection. These characteristics make them suitable as landmarks for robust matching when the cameras are moving around in an environment.

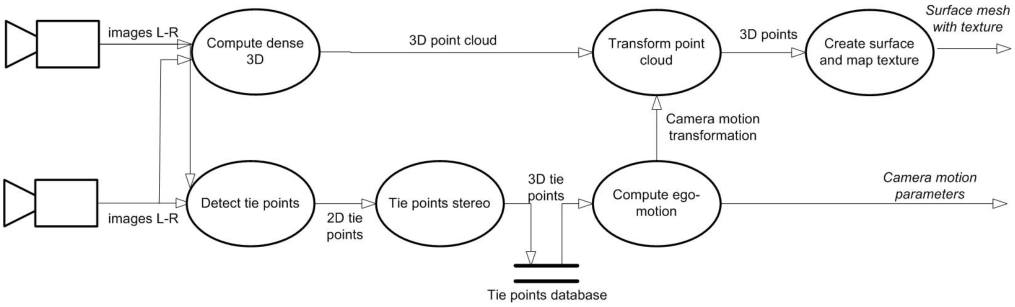


Figure 1. ***iSM system architecture.***

Previous approaches to feature detection, such as the widely used Harris corner detector [5], 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.

3.3. Camera Ego-motion Estimation

With 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 for each matched feature is also computed. Typically, we obtain hundreds of SIFT 3D features.

Subsequently we can compute the 3D position (X, Y, Z) of each stereo matched SIFT feature, using the following equations:

$$X = \frac{(u - u_0)I}{d}; \qquad Y = \frac{(v_0 - v)I}{d}; \qquad Z = \frac{fI}{d}$$

where (u, v, d) are the SIFT image location and disparity, (u_0, v_0) are the image centre coordinates, I is the baseline distance and f is the focal length.

For our hand-held camera, we recover the 6 dof camera ego-motion when the camera moves freely in 3D. We employ a Simultaneous Localization And Mapping (SLAM) approach that uses the SIFT features 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.* [10] reported a 27.7% reduction in rover navigation error when multi-frame tracking is used, rather than considering each pair of frames separately.

As the SIFT features are highly distinctive, they can be matched with very few false matches. We can then find the camera movement that would bring each projected SIFT feature into the best alignment with its matching feature. A weighted least squares procedure is carried out taking into account the feature uncertainty. Features with large least squares errors are discarded as outliers.

3.4. Mesh Creation

As the camera moves around, dense 3D data is obtained relative to the camera position at each frame. The initial camera pose is used as the reference and all 3D data sets are transformed to this coordinate system using the camera pose estimated for each data set.

Using all 3D points obtained from the stereo processing is not efficient as there is 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 [16], 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.

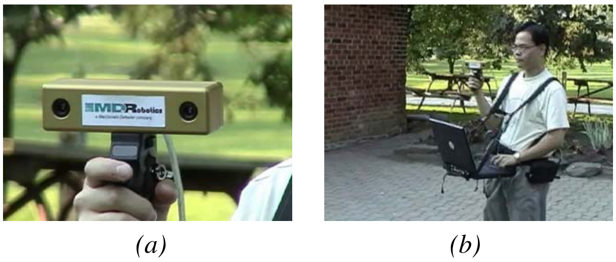
3.5. Texture Mapping

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 seamlines 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.

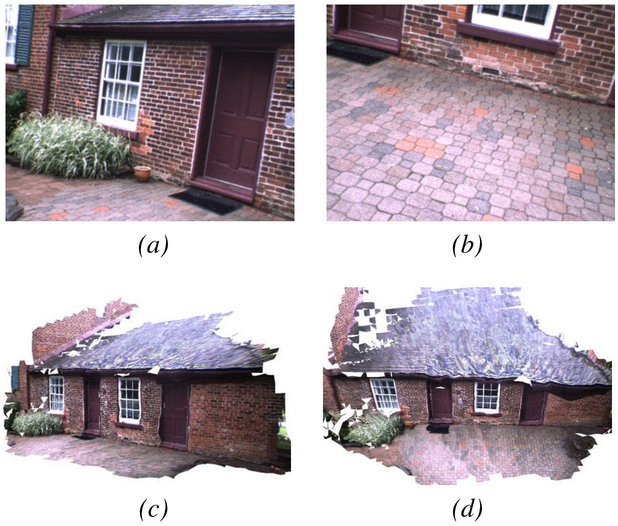
4. Experimental Results

Figure 2(a) shows the iSM hand-held stereo camera. We use a laptop PC with a Pentium IV 2.4GHz processor and 1GB RAM. As laptops only provide firewire interface without power, an external hub with power supply is required.



(a) *(b)*

Figure 2. ***(a) iSM hand-held stereo camera. (b) iSM laptop carried by the user on a harness.***



(a) *(b)*

(c) *(d)*

Figure 3. ***(a) and (b) Two input images from the house sequence. (c) and (d) Two views of the resultant 3D model.***

For portability, a battery is used to supply power to the hub and camera; the laptop is carried by the user on a harness, as shown in Figure 2(b).

In one of the experiments, we modeled a facade of a house. The camera was moved freely pointing at different portions of the house and about 30 seconds of images were captured. Figure 3(a) and (b) show two hand-held images from the house sequence. Then, the system processed these images automatically and created a photo-realistic 3D model in around 5 minutes.

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 3(c) and (d) show 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.

Dense stereo depth resolution is given by:

$$\Delta Z = \frac{\Delta d}{fI}\, Z^2$$

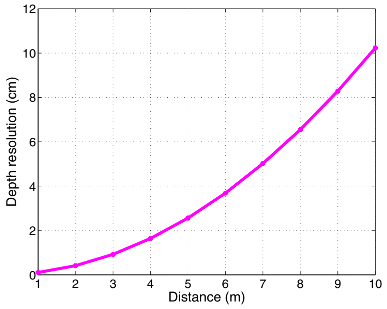


Figure 4. ***Stereo camera depth resolution at various distances.***

The depth resolution decreases quadratically with the distance and improves as the baseline or the image resolution increases. Figure 4 shows the depth resolution of our Bumblebee stereo camera at various distances, with 12cm baseline at 640x480 image resolution and 0.1 pixel disparity resolution (Δd).

Experimental results show that for scenes within 3m from the camera, our 3D models have an accuracy within 2cm, including dense stereo uncertainty and camera motion estimation error. Stereo camera with wider baseline and higher resolution can be used for larger environments to achieve better accuracy.

5. 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. This capability is useful for creating 3D model for larger environments, as it can automatically combine individual 3D models that cover smaller environments.

We will make use of the highly distinctive SIFT features to do the 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 [20], 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 is as follows:

- 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 [4] is given by:

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

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 $p = 3$, $\epsilon = 0.8$, $\tau = 99\%$, $m = 573$. 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.

Two separate image sequences have been captured in the lab and two 3D models are obtained with some overlapping region. 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.

Figure 5(a) and (b) show the two individual 3D models. Figure 5(c) 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 proposed SIFT-based auto-referencing algorithm can align two 3D models automatically and multiple 3D models can be aligned in an incremental or pair-wise fashion.

6. Pattern Projector

There is often not much texture in indoor man-made environments 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, as shown in Figure 6(a). The projector is synchronized with the stereo camera and the projector trigger is controlled via software.

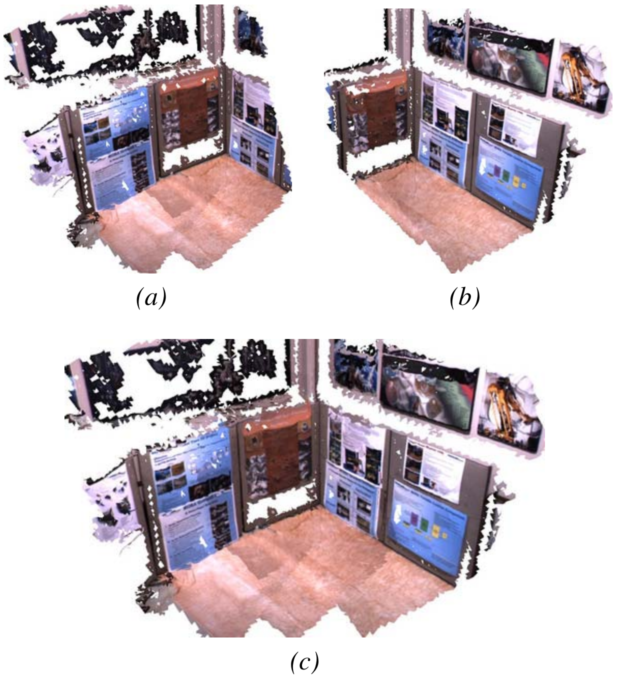


Figure 5. ***(a) and (b) Two 3D models with some overlap but unknown alignment. (c) Aligned model after auto-referencing.***

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 6(b) and (c) show an example of a normal image and a flash image showing the random dot pattern. Figure 6(d) and (e) show 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. We can see that the coverage of the 3D model is increased substantially with the pattern projector.

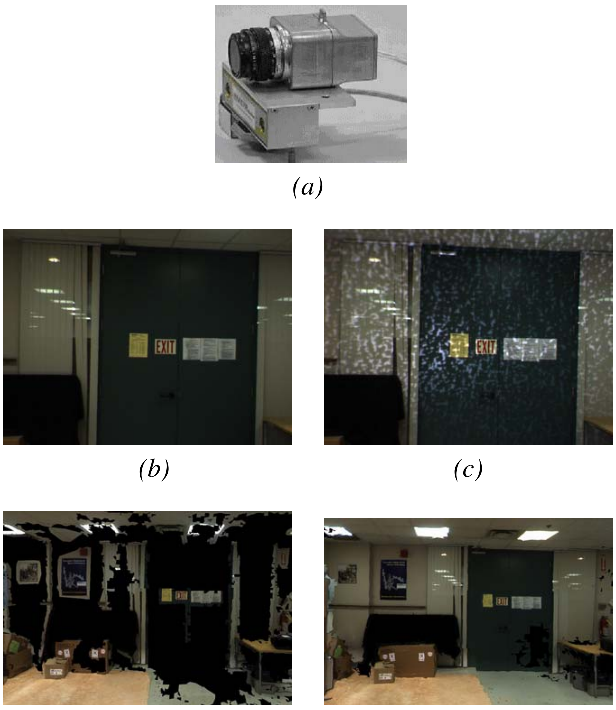
7. Applications

Forensics, mining, autonomous vehicles/planetary rovers are among the many applications that can make use of photo-realistic 3D calibrated models.

7.1. Forensics

Crime scene reconstruction is one of the important aspects of police and forensics investigation. Currently, the police spends a lot of time at each crime scene taking many photos and performing manual measurements.

With iSM, they can create a 3D model of the crime scene quickly without much disturbance to the crime scene and measurements can be performed on the 3D model after-



(a) *(b)* *(c)* *(d)* *(e)*

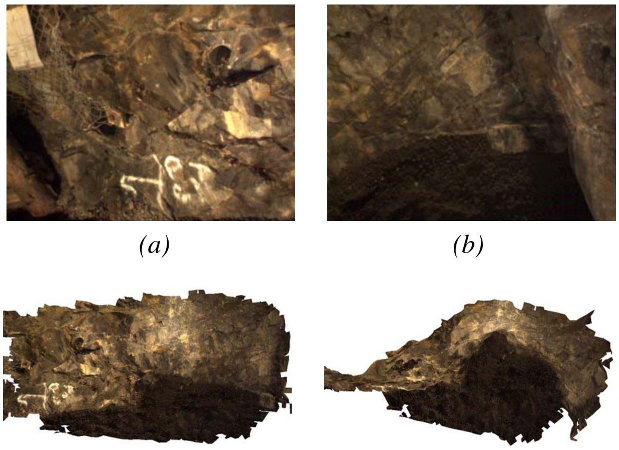
Figure 6. ***(a) Stereo camera with flash pattern projector. (b) A sample normal image. (c) A sample flash image. (d) A screenshot of a 3D model without the pattern projector. (e) A screenshot of a 3D model with the pattern projector.***

wards [18]. 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 also be shown in the court so that the judge and the jury can understand the crime scene better.

7.2. Mining

Photo-realistic 3D models are useful for surveys 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.

Currently, geological surveys are 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. Past and recent models can be linked together and used for planning the exploration and tunnel advancement.



(a) *(b)* *(c)* *(d)*

Figure 7. ***(a) and (b) Two input images from the underground mine cavity sequence. (c) and (d) Two views of the resultant 3D model.***

We have carried out field trials in several underground mines. Lights are attached to the camera to provide illumination. Figure 7(a) and (b) show two images from the hand-held images captured in an underground mine cavity. Mine images provide very rich features for both SIFT extraction and dense stereo matching. This image sequence was captured in around 20 seconds and the processing took less than 5 minutes. Figure 7(c) and (d) show two views of the 3D model generated from the input sequence. Subsequently, 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.

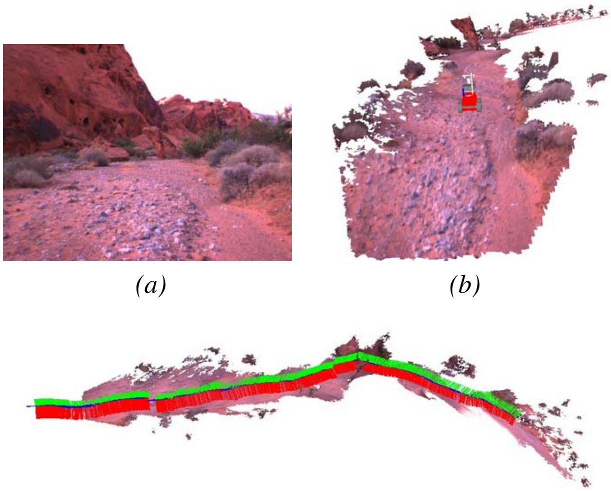
7.3. Autonomous Vehicles

For autonomous vehicles and planetary rovers, the creation of 3D terrain models of the environment is useful for visualization and path planning. Apart from the 3D model, iSM also computes the camera ego-motion which allows the vehicles to localize themselves more accurately, as wheel odometry is prone to slippages over time.

iSM has been tested on a rover traversing over 40 metres in a desert in Nevada. Figure 8(a) shows an image from a sequence captured by the rover. Figure 8(b) shows the reconstructed 3D terrain model with a virtual rover inserted for visualization, and Figure 8(c) shows the recovered camera trajectory without using wheel odometry.

8. Conclusion

In this paper, we have presented a 3D modeling system, the instant Scene Modeler (iSM). iSM uses a hand-held stereo camera for recording images and a laptop for acquisition and processing. It creates photo-realistic 3D calibrated models of environments automatically (no user interaction) within minutes. We also proposed the auto-referencing al-



(a) *(b)*

(c)

Figure 8. *(a) An image from a rover sequence taken in a desert in Nevada. (b) Reconstructed 3D model with rover model inserted. (c) Recovered camera path.*

gorithm which can align multiple 3D models automatically and some initial experimental results are shown. A random dot pattern projector has been integrated to obtain better 3D models in indoor man-made environments. We have also discussed the benefits of using iSM for crime scene modeling, mining and autonomous vehicles, and shown some promising results in these applications.

The quality and accuracy of the 3D model depend on the camera motion estimation and the dense stereo matching. Camera ego-motion accuracy is affected by the number of SIFT features in the environment and their distribution. Ego-motion error decreases when the number of SIFT features increases and when features are more widely spread out. However, long-term drifts in the motion estimation may occur for long sequences.

For future work, external sensors such as an orientation sensor can be incorporated to augment the visual motion estimation. Backward correction technique [20] can be applied to improve multiview registration and loop closure.

References

[1] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE PAMI*, 14(2):239–256, 1992.

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

[3] Cyberware. http://www.cyberware.com (Jan 12, 2006).

[4] 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.

[5] C.J. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of 4th Alvey Vision Conference*, pages 147–151, Manchester, 1988.

[6] P. Hebert. A self-referenced hand-held range sensor. In *Proceedings of 3DIM*, pages 5–12, Quebec City, Canada, 2001.

[7] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of ICCV*, pages 1150–1157, Kerkyra, Greece, September 1999.

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

[9] D. Nister. Automatic passive recovery of 3d from images and video. In *Proceedings of Second International Symposium on 3D Data Processing, Visualization & Transmission (3DPVT04)*, Thessaloniki, Greece, September 2004.

[10] C.F. Olson, L.H. Matthies, M. Schoppers, and M.W. Maimone. Robust stereo ego-motion for long distance navigation. In *Proceedings of CVPR Volume 2*, pages 453–458, South Carolina, June 2000.

[11] PhotoModeler. http://www.photomodeler.com (Jan 12, 2006).

[12] Polhemus. http://www.polhemus.com (Jan 12, 2006).

[13] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool. Hand-held acquisition of 3d models with a video camera. In *Proceedings of 3DIM*, pages 14–23, Ottawa, 1999.

[14] V. Popescu, E. Sacks, and G. Bahmutov. The modelcamera: a hand-held device for interactive modeling. In *Proceedings of 3DIM*, pages 285–292, Banff, Canada, October 2003.

[15] Point Grey Research. http://www.ptgrey.com (Jan 12, 2006).

[16] G. Roth and E. Wibowo. An efficient volumetric method for building closed triangular meshes from 3-d image and point data. In *Proceedings of Graphics Interface (GI)*, pages 173–180, Kelowna, B.C., Canada, 1997.

[17] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. In *Proceedings of SIGGRAPH'02*, San Antonio, Texas, July 2002.

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

[19] S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *IJRR*, 21(8):735–758, August 2002.

[20] S. Se, D.G. Lowe, and J.J. Little. Vision-based global localization and mapping for mobile robots. *IEEE Transactions on Robotics*, 21(3):364–375, June 2005.

[21] 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.

[22] G. Turk and M. Levoy. Zippered polygon meshes from range images. In *Proceedings of SIGGRAPH'94*, pages 311–318, Orlando, Florida, July 1994.