Proc. 2002 Australasian Conference on Robotics and Automation
Auckland, 27-29 November 2002

3D Vision for Large-Scale Outdoor Environments

Dorian J. Spero and Ray A. Jarvis
Intelligent Robotics Research Centre
Monash University,
Clayton, Victoria 3168, Australia
Email: dorian@ieee.org

Abstract

The difficulty of inferring depth from a 2D image has resulted in a growing interest in rangefinding technologies. This paper describes the development of a low-cost 3D vision system that was used to extract a rich data set from an unstructured outdoor domain. The system consists of a laser rangefinder (ladar) and monocular camera, both mounted on a pan-tilt unit for accurate positioning. By sensor fusion of ladar data with still images captured by the camera, a high-resolution environmental model was created that represents surface shape and colour. Preliminary results in proximity, surface normal directionality, and colour based scene segmentation through semantic-free clustering processes, are presented. In the context of robotics, a future project will utilise this new sensor system for the purpose of natural landmark based localisation in rough terrain.

1 Introduction

Anthropomorphically speaking, vision is the richest source of sensory data in support of rudimentary tasks such as navigation and object manipulation. This has resulted in a large research contingent in the area of computer vision. However, attempts to artificially mimic human vision have yet to reach the enormous expectations borne of our own facility.

The fundamental problem of obtaining depth from a 2D image, captured by a monocular camera, remains a difficult task. Consequently, the field of rangefinding has grown to be an integral part within the area of computer vision, in an endeavour to find complementary techniques that overcome the limitations of current camera technology. By fusing 2D vision with rangefinding sensors, as demonstrated in [Jarvis, 1992], a solution to 3D vision is realised - circumventing the problem of inferring 3D from 2D.

In an unstructured outdoor environment, extracting 3D shape and surface colour involves developing a 3D vision sensor that can accommodate large-scale space, while striking the appropriate balance between cost, safety, accuracy, robustness, weight and power consumption. It is widely known that passive rangefinding methods are generally limited to short range applications and their performance is highly dependent on ambient environmental conditions [Jarvis, 1993]. These factors impede their usefulness in large-scale outdoor space.

Figure 1: Experimental Rig

As a representative example, lateral stereopsis is a commonly used passive rangefinding method that involves finding the disparity in image position of a discernible entity viewed by multiple, geometrically spaced, cameras. Range measurements are calculated using triangulation, parameterised by the disparity, baseline length and vergence between cameras. With a baseline length of the same separation as human eyes, lateral stereopsis is a viable depth cue modality for a range of only several metres. Increasing the baseline simultaneously improves the depth accuracy and reduces the view all cameras share. For computer vi-

Copyright © ARAA 2002

sion it is possible to increase the baseline considerably for long-range outdoor work but rigid alignment of the cameras becomes a problem. In any case, the range accuracy remains inversely proportional to actual depth due to the triangulation geometry. Time-of-flight laser ranging, an active rangefinding method, is not subject to this constraint and can be reliably used for long-range applications (as in [Guivant *et al.*, 2000; Mayora *et al.*, 1998]).

In the present investigation, the 3D vision system consists of a laser rangefinder (ladar) and monocular camera, both mounted on a pan-tilt unit for accurate positioning. Figure 1 depicts the experimental rig that was used in an outdoor environment. Ladar data was gathered and fused with still images captured by the camera, producing a rich data set consisting of x, y, z triplets, in 3D Cartesian space, and associated RGB intensity values. Semantic-free clustering analysis was performed, for the purpose of segmenting the raw data set in a manner suitable for generalised high-level interpretation.

In section 2, the sensor system is described. The sensor fusion process is detailed in section 3, followed by semantic-free analysis in section 4. The paper concludes with a discussion and future work in section 5. To prove the usefulness of the proposed 3D vision sensor, experimental results are laced throughout.

2 Sensor System

Figure 2 shows the sensing system, which is composed of a ladar (ILM300HR from Measurement Devices), monocular camera (Webcam 5 from Creative) and pan-tilt unit (PTU-46-17.5 from Directed Perception). The camera is directly attached to the ladar, via an aluminium bracket, for the purpose of aligning the optical axes of both sensors. This arrangement makes both sensors point in the same direction with a relatively small translation separating the ladar's optical center from the camera's principle point. The pan-tilt unit positions both sensors, in 3D space, simultaneously.

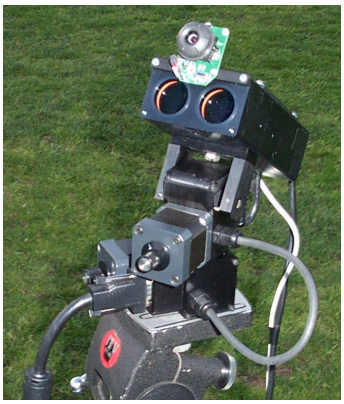

Figure 2: Camera, Ladar and Pan-Tilt Unit

The ladar was used to extract the scene geometry and since it is an active ranging system it can operate irrespective of ambient lighting conditions. The 905nm wavelength ladar has a maximum range of 300m, a typical accuracy of 30cm and makes a measurement by using time-of-flight of a single laser pulse (single shot mode) at a repetition rate of 1000Hz.

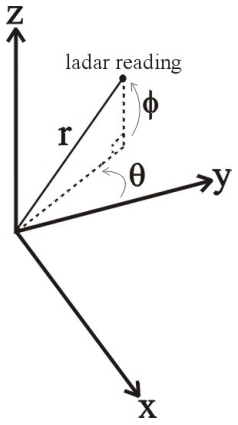

Figure 3: Ladar Reference System

Range readings were taken in polar coordinates as $r = f(\theta, \phi)$, based on the reference system given in Figure 3. The pan-tilt unit provides a 318° horizontal vision field ($\theta \in [-159°, 159°]$) and a 77.7° vertical vision field ($\phi \in [-46.7°, 31°]$). The horizontal and vertical angular resolution is 0.0514°. Range readings were converted from polar to Cartesian coordinates, $z = f(x, y)$, using the transformation in Equation 1.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -r\sin(\theta)\cos(\phi) \\ r\cos(\theta)\cos(\phi) \\ r\sin(\phi) \end{bmatrix} \quad (1)$$

In a complementary fashion to the volumetric data obtained by the ladar, the camera was used to extract the corresponding intensity data. The 24-bit colour camera has a resolution of 640x480 pixels and is capable of capturing 15 frames per second. Parameters such as the focal length and field of view, were not supplied by the manufacturer and therefore had to be obtained through calibration (as discussed in Section 3.3).

Both sensors are positioned or slewed, in unison, by adjusting the azimuth (θ) and elevation (ϕ) of the pan-tilt unit. Having the sensors in close proximity with their optical axes aligned, means that the camera can be positioned such that its field of view encompasses an arbitrary range reading gathered from the ladar. The sensor proximity also simplifies the sensor fusion problem, as affine transformations (translations and rotations) between sensor data sets are essentially eliminated with this virtual single unit arrangement.

3 Sensor Fusion

To extract the 3D shape and surface colour of an outdoor field, two independent data sets were gathered sequentially. A data set of x, y, z triplets and a data set of 24-bit RGB intensity values, were obtained via a ladar scan and a vision scan, respectively.

3.1 Ladar Scan

A ladar scan consists of taking ladar measurements at varying azimuth and elevation angles. In the present investigation, for timely 3D data acquisition, the pan-tilt unit was set to a particular elevation angle and then horizontally slewed, while concurrently matching the timing of the ladar readings to the adjustable slew speed profile of the pan-tilt unit. By incrementally adjusting the elevation angle and slewing the ladar in a back and forth manner, dense volumetric data was obtained. The data was stored in a 2D array, indexed by the azimuth and elevation angles. Invalid data (that is, ranges exceeding the ladar's capabilities) were zeroed and then stored, to maintain proper order within the array.

Figure 4 shows a ladar scan that was performed in an outdoor environment. The scan covered a horizontal range of $[-154.3°, 154.3°]$ and a vertical range of $[-25.7°, 20.6°]$. The horizontal angular resolution was set to $0.0514°$, with the vertical angular resolution set to $0.257°$. This resulted in approximately 1.1 million data points being gathered over a time period of 18 minutes. To pictorially indicate variations in depth (z), a colour ramp has been used, whereby a colour gradient from red to blue represents the depth range from low to high, respectively.

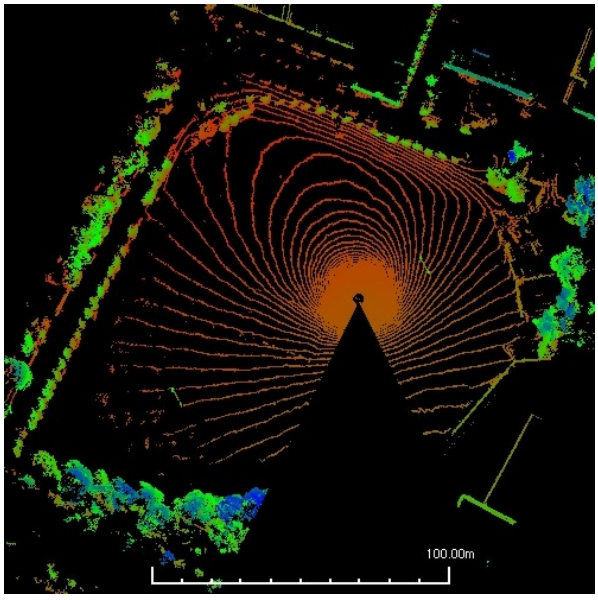

Figure 4: Ladar Scan of an Outdoor Environment

3.2 Vision Scan

Similar to a ladar scan, a vision scan consists of capturing still images at varying azimuth and elevation angles. Each still image was captured by first positioning the pan-tilt unit and then grabbing a camera image after a short time delay, dependent on the camera's settling time and shutter speed. The number of still images taken and their corresponding positions, were based on satisfying a criteria, including: encompassing every ladar reading within the effective field of view yielded by the image set; minimising the effects of lens distortion and adhering to memory limitations. Consequently, every x, y, z triplet, from the ladar scan, had an associated image pixel.

The camera was found to have a pronounced barrel distortion, which is defined as a negative radial displacement of the image points. Outer points tend to crowd increasingly together, along with a decrease in scale [Weng *et al.*, 1992]. The effects of lens distortion was minimised by visually overlapping, to a partial extent, each image with its neighbouring images, thus allowing the distorted outer edges to be excluded from the sensor fusion process.

Figure 5 depicts a partial image sequence from a vision scan, performed to complement the ladar scan described in Section 3.1. The scan consisted of 20 still images, with two rows of 10 images taken at an elevation angle of $9°$ and $-14.1°$, respectively. Images within each row were taken from an azimuth angle of $-138.9°$ to $139.9°$, at intervals of $30.0°$. The images were stored in 2D arrays, each supplemented with position data.

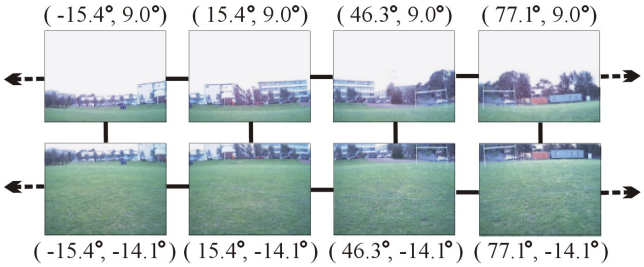

Figure 5: Partial Image Sequence

3.3 Volumetric and Intensity Data Registration

An image-based technique was devised for registering the volumetric and intensity data, gathered from the scans described in Sections 3.1 and 3.2, respectively. The technique directly matches image pixels to ladar points, by mapping the intensity data arrays to the volumetric data array via linear interpolation.

Figure 6 shows the registration process, where one image array chosen amongst the 20 images in the image set, is mapped to the volumetric data array. The process is based on a reference system that consists of

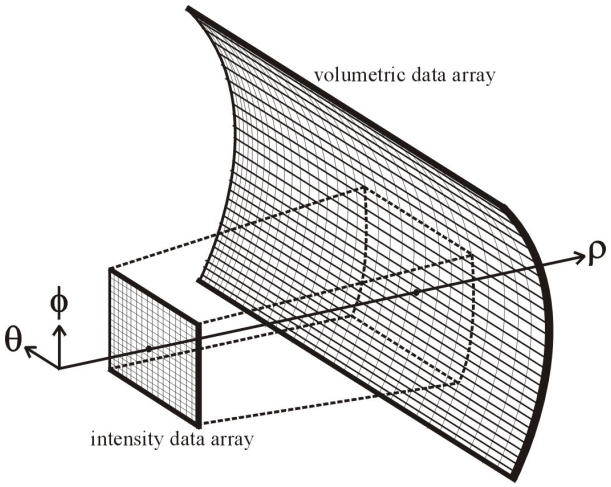

Figure 6: Registration Process

azimuth (θ), elevation (ϕ) and depth (ρ) dimensions, fixed at the centroid location of each image. The image and volumetric data arrays are stored with the same orientation in memory, simplifying the indexing task during each intensity to range datum match.

To register a ladar range with an intensity value, the location (θ, ϕ) and range (r) of the specific volumetric data element is first obtained and then the most appropriate image selected. Image selection is based on finding the closest image, given by its centroid position (θ_0, ϕ_0), to the aforementioned ladar location (θ, ϕ). This means that registration will be performed using an image whose optical center is closest, thus minimising the effects of lens distortion. Given that (u_i, v_j) is the set of image pixels from the selected image, the indexes u and v were found using Equations 2 and 3, respectively. The origin (u_0, v_0) is located at (θ_0, ϕ_0), with constraints $-\frac{R_u}{2} \preceq i \prec \frac{R_u}{2}$ and $-\frac{R_v}{2} \preceq j \prec \frac{R_v}{2}$, where R_u is the horizontal image resolution (640 pixels) and R_v is the vertical image resolution (480 pixels).

$$u = \frac{R_u}{V_\theta}(\theta - \theta_0)(1 - \frac{r\cos(\theta - \theta_0)}{g_\rho}) \tag{2}$$

$$v = \frac{R_v}{V_\phi}(\phi - \phi_0)(1 - \frac{r\cos(\phi - \phi_0)}{h_\rho}) \tag{3}$$

The parameters V_θ (horizontal field of view), V_ϕ (vertical field of view), g_ρ (horizontal depth of field) and h_ρ (vertical depth of field), were found experimentally using the calibration tool depicted in Figure 7. Assuming that each pixel has sides of equal length (that is, square shaped), only one field of view parameter needed to be estimated. The angle subtended by the other parameter is related via R_u and R_v.

To convert from the camera's perspective projection to orthographic projection, the g_ρ and h_ρ parameters were used to place a vanishing point on the camera's optical axis. The further the depth of a ladar point, acquired from (r, θ, ϕ), the closer the point was pulled to the optical axis. In this case, the $\frac{1}{\rho}$ variation in object size was represented by a linear function, with g_ρ and h_ρ inversely proportional to the contraction rate.

Figure 7: Calibration Tool

The calibration tool (Figure 7) is a board composed of a colourful rectangular pattern. Rectangular perforations were purposefully bored into the tool, so the ladar would return a recognisable formation. By placing the tool at multiple locations in the operating environment and performing ladar and vision scans, the four unknown parameters were obtained through visually matching the ladar pattern to an intensity image. Once the camera was calibrated, with respect to the ladar, sensor fusion became an automated process - generating the complete set of $(x, y, z, red, green, blue)$ sextuplets.

Figure 8 shows the sensor fusion results. Each element from the volumetric data array, in sequence, was registered with an image pixel. The resulting data set was rendered using OpenGL, creating a realistic model of the 3D shape and surface colour of objects within the outdoor scene. Gaps in the model were partially filled by recursively interpolating new range values from the existing set. Several features are clearly identifiable, such as tree formations, buildings and a grassy oval.

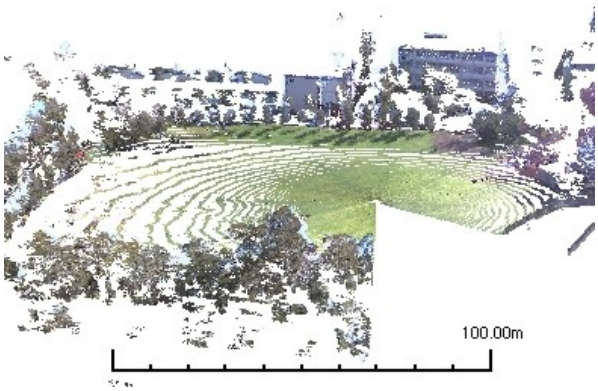

Figure 8: Ladar and Vision Sensor Fusion

4 Semantic-Free Analysis

The rich data set supports a wide variety of semantic-free clustering for the purposes of revealing useful structure. Segmenting a scene to identify features, and their corresponding properties, is pertinent to applications such as mobile robot localisation [Borenstein *et al.*, 1997], robot manipulation [Jarvis, 1992] and target tracking [Maimone *et al.*, 1999]. However, segmenting a data set of such high density poses a computational problem, when accessing each data element and its neighbours for comparative analysis. Consequently, by disregarding a portion of the available data set, a lower resolution was adopted for timely response.

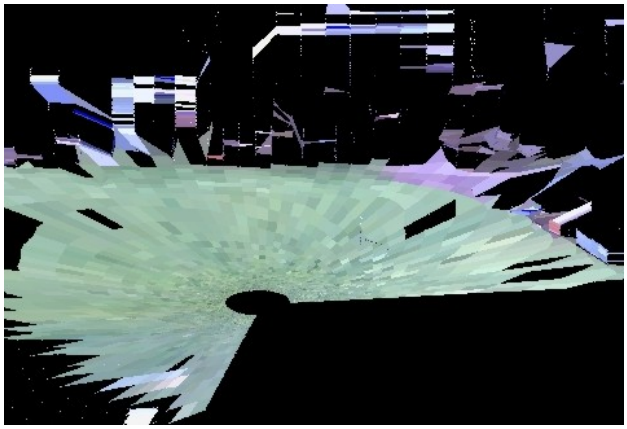

Figure 9: Colour Image Segmentation

Figure 9 shows colour based image segmentation, where homogeneously coloured pixels are grouped. The clusters were formed by thresholding the difference between the RGB intensity values of each data point and its neighbours. Conversely, Figure 10 shows scene segmentation based on spatial geometry. Lighter regions depict clusters that have data points in close proximity, determined by thresholding in Cartesian space.

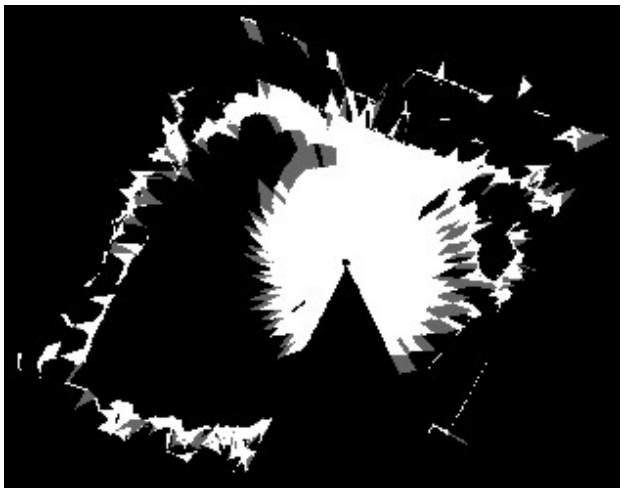

Figure 10: 3D Space Geometry Clustering

Unit surface normal components (n_x, n_y, n_z) were incorporated into the data set, for clustering based on surface slope and directionality. Since the volumetric data array is stored uniformly in memory, the immediate neighbourhood of data points simplifies triangulation and the construction of data meshes. Unit normals were generated (Figure 11), by calculating the normalised cross product between two vectors from a data point to its neighbours in the mesh. Figure 12 shows clustering based on surface normal directionality. Flat faces of the buildings and terrain are identifiable, along with sporadically occurring chaotic directionality indicating the possible presence of trees or bushes.

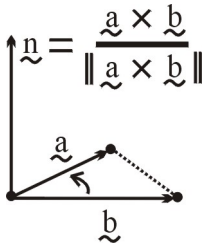

Figure 11: Unit Normal Generation

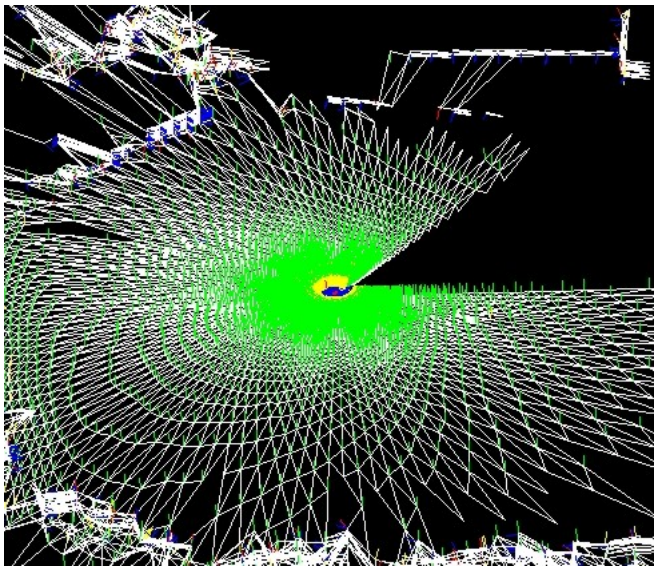

Figure 12: Surface Normal Directionality

The final data set was extensively featured, with vectors consisting of nine attributes:

- x, y, z - Cartesian coordinates
- r, g, b - colour intensity components
- n_x, n_y, n_z - unit surface normal components

A subset of these attributes was used to extract desirable clusters. However, the combination of all attributes, in some purposeful manner, can be used to form clusters that satisfy highly restrictive properties.

The resulting flexibility, afforded to low-level clustering analysis, assists with the development of a robust scene segmentation method to suit specific application needs.

5 Discussion and Conclusions

The presented sensor system was found to be a relatively low-cost solution to 3D vision. The system produced accurate, high-resolution scans of an outdoor domain, proving its applicability to large-scale areas. Having the sensors in such close proximity, simplified the volumetric and intensity data registration. Unlike the sensor arrangement described in [Cobzas *et al.*, 2002], a full 3D rigid transformation between ladar and camera reference systems, was not needed. In terms of efficiency, the time period taken for a complete environmental scan is not suited for real-time tasks. However, the scan time can be significantly reduced by focusing the scan (that is, decreasing the field of view), decreasing the scan resolution or increasing the scan slew rate.

The Universal Serial Port (USB) interface of the camera enhanced system portability, as numerous video cameras require hardware installation. For instance, the video cameras that require a frame grabber card to be connected to the Peripheral Component Interface (PCI) bus on a computer's motherboard, prevents the use of most portable computers such as laptops, notebooks and Microspace PCs (compact PCs from Digital Logic). In support of off-board processing for mobile robots, especially those with severe size and/or weight restrictions, development of a network controlled turnkey for the 3D vision system is underway. The system will ultimately be controlled via wireless Ethernet, further enhancing system portability.

Preliminary results in semantic-free analysis showed that useful clusters can be extracted from the rich data set, based on a multitude of segmentation options over the feature vectors. Semantic-free analysis was found to be a robust methodology, but the inclusion of domain specific analysis is still acknowledged as an important step for system efficiency. Future work will include developing a clustering technique, similar to [Jarvis and Patrick, 1973], for the versatile removal of data outliers during the segmentation process.

It is intended that the 3D vision system be incorporated into an existing robot path planner, devised in [Spero and Jarvis, 2002], and used for a future project focusing on natural landmark based localisation ([Borenstein *et al.*, 1997]) in rough terrain. In an attempt to counteract extreme lighting conditions inherent in an outdoor environment (for example, the sun's glare), a light filter will be sourced for the camera. This will restrict its dynamic operating range, thus maintaining data integrity and improving the robustness of the system in outdoor environments.

References

[Borenstein *et al.*, 1997] J. Borenstein, H. R. Everett, L. Feng, and D. Wehe. Mobile robot positioning: Sensors and techniques. *Journal of Robotic Systems*, 14(4):231–249, Apr. 1997.

[Cobzas *et al.*, 2002] D. Cobzas, H. Zhang, and M. Jagersand. A comparative analysis of geometric and image-based volumetric and intensity data registration algorithms. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, volume 3, pages 2506–2511, Washington, DC, 11-15 May 2002.

[Guivant *et al.*, 2000] J. Guivant, E. Nebot, and S. Baiker. High accuracy navigation using laser range sensors in outdoor applications. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation*, volume 4, pages 3817–3822, San Francisco, CA, 24-28 Apr. 2000.

[Jarvis and Patrick, 1973] R. A. Jarvis and E. A. Patrick. Clustering using a similarity measure based on shared near neighbours. *IEEE Transactions on Computers*, C-22(11):1025–1034, Nov. 1973.

[Jarvis, 1992] R. A. Jarvis. 3D shape and surface colour sensor fusion for robot vision. *Robotica*, 10:389–396, 1992.

[Jarvis, 1993] R. Jarvis. *Three-Dimensional Object Recognition Systems*, chapter Range Sensing for Computer Vision. Elsevier, Singapore, 1993.

[Maimone *et al.*, 1999] M. W. Maimone, I. A. Nesnas, and H. Das. Autonomous rock tracking and acquisition from a mars rover. In *Fifth International Symposium on Artificial Intelligence, Robotics and Automation in Space*, pages 329–334, Noordwijk, Netherlands, 1-3 Jun. 1999.

[Mayora *et al.*, 1998] K. Mayora, I. Moreno, and G. Obieta. Perception system for navigation in a 3D outdoor environment. In *Proceedings of IAV'98 3rd IFAC Intelligent Autonomous Vehicles*, pages 381–385, Madrid, Spain, 25-27 Mar. 1998.

[Spero and Jarvis, 2002] D. J. Spero and R. A. Jarvis. Path planning for a mobile robot in a rough terrain environment. In *Third International Workshop on Robot Motion and Control*, Bukowy Dworek, Poland, 9-11 Nov. 2002. IEEE. In Press.

[Weng *et al.*, 1992] J. Weng, P. Cohen, and M. Herniou. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(10):965–980, Oct. 1992.