# Inertial Sensed Ego-motion for 3D Vision

**Jorge Lobo and Jorge Dias**
*Institute of Systems and Robotics*
*DEEC, University of Coimbra-Polo II*
*3030-290 Coimbra, Portugal*
*e-mail: jlobo@isr.uc.pt, jorge@isr.uc.pt*

Received 4 November 2003; accepted 4 November 2003

Inertial sensors attached to a camera can provide valuable data about camera pose and movement. In biological vision systems, inertial cues provided by the vestibular system are fused with vision at an early processing stage. In this article we set a framework for the combination of these two sensing modalities. Cameras can be seen as ray direction measuring devices, and in the case of stereo vision, depth along the ray can also be computed. The ego-motion can be sensed by the inertial sensors, but there are limitations determined by the sensor noise level. Keeping track of the vertical direction is required, so that gravity acceleration can be compensated for, and provides a valuable spatial reference. Results are shown of stereo depth map alignment using the vertical reference. The depth map points are mapped to a vertically aligned world frame of reference. In order to detect the ground plane, a histogram is performed for the different heights. Taking the ground plane as a reference plane for the acquired maps, the fusion of multiple maps reduces to a 2D translation and rotation problem. The dynamic inertial cues can be used as a first approximation for this transformation, allowing a fast depth map registration method. They also provide an image independent location of the image focus of expansion and center of rotation useful during visual based navigation tasks.
© 2004 Wiley Periodicals, Inc.

## 1. INTRODUCTION

Biological vision systems are known to incorporate other sensing modalities. The inner ear vestibular system in humans and in animals provides inertial sensing mainly for orientation, navigation, control of body posture, and equilibrium. This sensorial system also plays a key role in several visual tasks and head stabilization, such as gaze holding and tracking visual movements.[1] Neural interactions of human vision and vestibular system occur at a very early processing stage.[2]

Artificial vision systems can provide better perception of their environment by using inertial sensor measurements of camera pose (rotation and translation). As in human vision, low level image processing

should take into account the ego motion of the observer. Nowadays micromachined low cost inertial sensors can be easily incorporated in computer vision systems, providing an artificial vestibular system. Inertial sensing is totally self-contained, except for gravity which provides an external reference.

This work is part of ongoing research into the fusion of inertial sensor data in computer vision systems. In ref. 3 a framework is set for vision and inertial sensor cooperation. The use of gravity as a vertical reference is explored, enabling camera focal distance calibration with a single vanishing point, vertical line segmentation, and ground plane segmentation. In ref. 4 world vertical feature detection and 3D mapping is presented. In this article we continue to explore the use of inertial data in vision systems, and present a method for fast alignment and segmentation of depth maps obtained from correlation based stereovision.

### 1.1. Related Work

Navigation in aerospace and naval applications has long relied on high grade inertial sensors.[5,6] The electronic and silicon micromachining development has produced low cost, batch fabricated, silicon sensors. Currently they are not suitable for stand-alone inertial systems, but can be useful in many applications. The level of integration is increasing, and single chip inertial systems for inertial aided GPS navigation systems are being developed.[7] This development has enabled many new applications for inertial sensors, not just in robotics and computer vision but also in large consumer commercial devices, such as video camera vibration compensation.

In computer vision applications, Viéville and Faugeras have proposed the use of inertial sensors[8] and studied the cooperation of the inertial and visual systems in mobile robot navigation by using the vertical cue, rectifying images and improving self-motion estimation for 3D structure reconstruction.[9–12] Inertial sensors were used to improve optical flow for obstacle detection by Bhanu et al.;[13] inertial sensed ego motion compensation improved interest point selection, matching of the interest points, and the subsequent motion detection, tracking, and obstacle detection.

Comparison of camera rotation estimate given by image optical flow with output from a low cost gyroscope was done by Panerai and Sandini for gaze stabilization of a rotating camera.[14,15] In ref. 16 they also studied the integration of inertial and visual information in binocular vision systems.

Mukai and Ohnishi used a gyroscope sensor to discriminate rotation and translation effects on the image and improve the accuracy of 3D shape recovery.[17,18] In ref. 19 Kurazume and Hirose used inertial sensors for image stabilization and attitude estimation of remote legged robots.

Virtual reality modelling and augmented reality are strong applications for inertial aided vision systems. Coorg et al.[20] use mosaicing and other techniques to perform an automated three-dimensional modeling of urban environments using *pose imagery* (i.e., images with known orientation and position obtained by inertial sensors and GPS). A hybrid inertial and vision tracking algorithm for augmented reality registration was proposed by Suya You et al.[21] Hoff et al. used a head mount system with inertial sensors and cameras, providing 3-D motion and structure estimation for augmented reality.[22,23]

A vision system for automated vehicles built by Dickmanns et al. has also incorporated inertial sensors.[24] The vision feature trackers use feedback from the inertial estimated state that has negligible time delays, and includes perturbations which must be taken into account by the vision system.

## 2. DATA FROM CAMERA SENSOR

Cameras can be seen as ray direction measuring devices. The pinhole camera model considers one center of projection, where all rays originated from world points converge. The image will be equivalent to a plane cutting that pencil of rays, projecting images of world points onto a plane. If we consider a unit sphere around the optical center, we can model the images as being formed on its surface. The image plane can be seen as a plane tangent to a sphere of radius $f$, the camera's focal distance, concentric with the unit sphere, as shown in Figure 1. Using the unit sphere gives an interesting model for central perspective and provides an intuitive visualization of projective geometry.[25,26] It also has numerical advantages when considering points at infinity, such as vanishing points.

### 2.1. Image Points

A world point $\mathbf{P}_i$ will project on the image plane as $\mathbf{p}_i$ and can be represented by the unit vector $\mathbf{m}_i$ placed at the sphere's center, the optical center of the camera. With image centered coordinates $\mathbf{p}_i = (x_i, y_i)$ we have

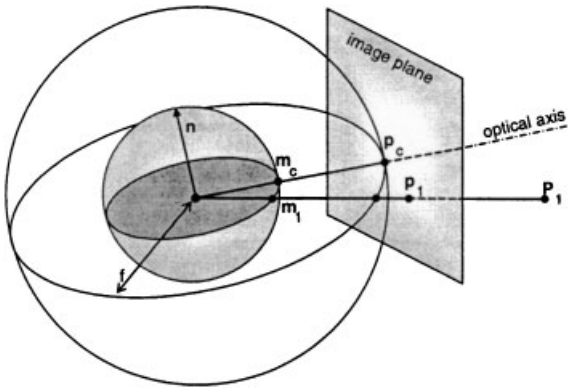


**Figure 1.** Point projection onto unit sphere.

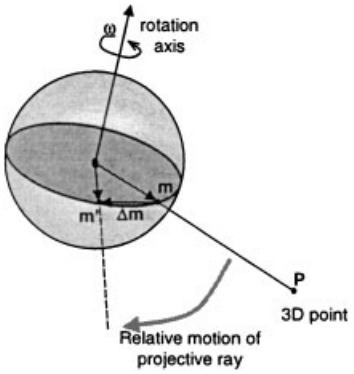


**Figure 2.** Projected unit sphere point motion with camera pure rotation.

$$\mathbf{P}_i \rightarrow \mathbf{m}_i = \frac{\mathbf{P}_i}{\|\mathbf{P}_i\|} = \frac{1}{\sqrt{x_i^2 + y_i^2 + f^2}} \begin{bmatrix} x_i \\ y_i \\ f \end{bmatrix}. \tag{1}$$

To avoid ambiguity $m_i$ is forced to be positive, so that only points on the image side hemisphere are considered.

### 2.2. Image Lines

Image lines can also be represented in a similar way. Any image line defines a plane with the center of projection, as seen in Figure 1. A vector $n$ normal to this plane uniquely defines the image line and can be used to represent the line. For a given image line $ax + by + c = 0$, the unit vector is given by

$$\mathbf{n} = \frac{1}{\sqrt{a^2 + b^2 + (c/f)^2}} \begin{bmatrix} a \\ b \\ c/f \end{bmatrix}. \tag{2}$$

As seen in Figure 1, we can write the unit vector of an image line with points $\mathbf{m}_1$ and $\mathbf{m}_2$ as

$$\mathbf{n} = \mathbf{m}_1 \times \mathbf{m}_2 . \tag{3}$$

### 2.3. Vanishing Points

Since the perspective projection maps a 3D world onto a plane or planar surface, phenomena that only occurs *at infinity* will project to very finite locations in the image. Parallel lines only meet at infinity, but in the image plane, the point where they meet can be quite visible and is called the *vanishing point* of that set of parallel lines.

A space line with the orientation of an unit vector $\mathbf{m}$ has, when projected, a *vanishing point* with unit sphere vector $\pm\mathbf{m}$. The vanishing point of a set of 3D parallel lines with image lines $\mathbf{n}_1$ and $\mathbf{n}_2$ is given by

$$\mathbf{m} = \mathbf{n}_1 \times \mathbf{n}_2 . \tag{4}$$

### 2.4. Ego-motion and Spherical Motion Field

When the camera sensor moves relative to the observed scene, image features will have a corresponding motion across the image. Using a spherical model, data from different camera configurations, such as omnidirectional images from catadioptric mirrors or several cameras with a common center of projection, can be incorporated into a unified model, with better spatial observability.

If the camera experiences a pure rotation $\boldsymbol{\omega}$, the fixed world $\mathbf{P}_i$ given in the camera referential will have a motion vector given by

$$\dot{\mathbf{P}}_i = -\boldsymbol{\omega} \times \mathbf{P}_i \tag{5}$$

as shown in Figure 2. The world point after the rotation $\mathbf{P}_i'$ is given by $\mathbf{P}_i - \boldsymbol{\omega} \times \mathbf{P}_i$. The unit sphere point after the rotation $\mathbf{m}_i'$ is given by

$$\mathbf{m}_i' = \frac{\mathbf{P}_i - \boldsymbol{\omega} \times \mathbf{P}_i}{\|\mathbf{P}_i - \boldsymbol{\omega} \times \mathbf{P}_i\|} = \mathbf{m}_i - \boldsymbol{\omega} \times \mathbf{m}_i . \tag{6}$$

Since the rotation is centered in the camera projective center, the induced image motion does not depend on the 3D point depth.

If the camera experiences both rotation $\boldsymbol{\omega}$ and translation $\mathbf{t}$ the fixed world $\mathbf{P}_i$ given in the camera referential will have a motion vector given by

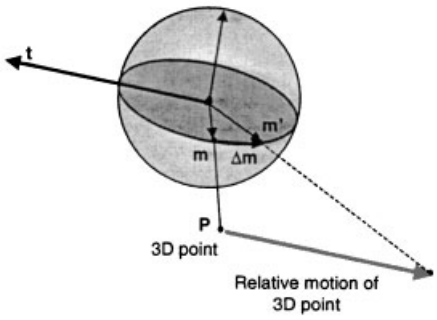


**Figure 3.** Projected unit sphere point motion with camera translation.

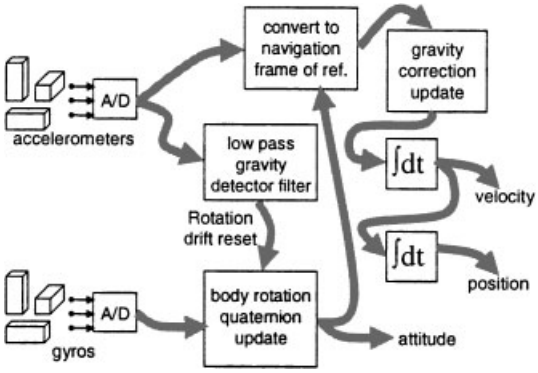


**Figure 4.** Simplified strapdown inertial navigation system.

$$\dot{\mathbf{P}}_i = -\mathbf{t} - \boldsymbol{\omega} \times \mathbf{P}_i \tag{7}$$

as shown in Figure 3. Projecting onto the unit sphere as before, the motion field on the unit sphere $\dot{\mathbf{m}}_i$ is given by

$$\dot{\mathbf{m}}_i = \frac{1}{\|\mathbf{P}_i\|}((\mathbf{t}.\mathbf{m}_i)\mathbf{m}_i - \mathbf{t}) - \boldsymbol{\omega} \times \mathbf{m}_i . \tag{8}$$

This equation describes the velocity vector $\dot{\mathbf{m}}_i$ for a given unit sphere point $\mathbf{m}_i$ as a function of camera ego-motion $(\mathbf{t},\boldsymbol{\omega})$ and depth $\|\mathbf{P}_i\|$.

## 3. DATA FROM INERTIAL SENSORS

At the most basic level, an inertial system simply performs a double integration of sensed acceleration over time to estimate position. But if body rotations occur, they must be taken into account. The measured accelerations are given in the body frame of reference, initially aligned with the navigation frame of reference. In strapdown systems the gyros measure the body rotation rate, and the sensed accelerations are computationally converted to the navigation frame of reference. Figure 4 shows a block diagram of a strapdown inertial navigation system. The system has an inertial measurement unit (IMU) with 3D orthogonal sets of acceleromters and gyrometers. Table I summarizes the data that can be obtained from the inertial sensors.

High grade sensors are required for inertial navigation, and low-cost MEMs inertial sensors offer low performance. Some assumptions can be made on the systems's dynamics to cope with the accumulated drift. If the norm of the sensed acceleration is about 9.8 $m.s^{-2}$, then we can assume that the accelerometers only measure $\mathbf{g}$, and the attitude can be directly determined, resetting the accumulated drift in the attitude computation. A low threshold can also be applied to the system, assuming that the system never accelerates or rotates below a certain value, preventing the error build up.

It is interesting to notice that human inertial sensing has a similar performance to currently available low-cost inertial sensors. Measuring the actual vestibular perceptual thresholds is difficult; they are determined by many factors such as mental concentration, fatigue, other stimulus capturing the attention, and vary from person to person.[27] Reasonable threshold values for perception of rotation are 0.14, 0.5 and 0.5 $deg.s^{-2}$ for yaw, roll, and pitch motions, respectively. Values of 0.01 g for vertical and 0.006 g for horizontal acceleration are appropriate representative thresholds for perceptible intensity of linear acceleration. These are valid for sustained and relatively low frequency stimulus.

The currently available low cost inertial sensors are capable of similar performances.[28] The inertial system prototype built for this work, using low cost sensors, has gyros with 0.1 $deg.s^{-1}$ resolution, and

**Table I.** Data from inertial sensors.

| | | |
|---|---|---|
| $\frac{d}{dt}$ | angular acceleration | $\varphi = \theta$ |
| | rate of linear acceleration (jerk) | $j = \dot{a} = \ddot{x}$ |
| | angular velocity | $\omega = \dot{\theta}$ |
| | linear acceleration + gravity | $a + g = \ddot{x} + g$ |
| $\int dt$ | angular position (attitude) | $\theta$ |
| | linear velocity | $v = \dot{x}$ |
| $\int\int dt$ | position | $x$ |

accelerometers with 0.005 g resolution. Notice that the gyros measure angular velocity and not angular acceleration.

These performances are not suitable for standalone inertial navigation, but combined with vision cues they contribute to human spatial orientation and body equilibrium. The inertial cues enhance the performance of the vision system in gaze stabilization, tracking, and visual navigation.

## 4. COMBINING STATIC INERTIAL CUES WITH VISION

### 4.1. Vertical Reference from Gravity

The measurements taken by the inertial unit's accelerometers include the sensed gravity vector summed with the body's acceleration. When the system is motionless, or subject to constant speed, gravity provides a vertical reference for the camera system frame of reference given by

$$\hat{\mathbf{n}} = \frac{\mathbf{a}}{\|\mathbf{a}\|}, \tag{9}$$

where $\mathbf{a}$ is the sensed acceleration, in this case the reactive (upward) force to gravity. By performing the rotation update using the IMU gyro data, gravity can be separated from the sensed acceleration. In this case $\hat{\mathbf{n}}$ is given by the rotation update, but must be monitored using the low-pass filtered accelerometer signals, for which the above equation still holds, to reset the accumulated drift.

The vertical unit vector is given in the IMU referential, and has to be converted to the camera referential. Only the rotation is relevant, and can be calibrated as described below.

### 4.2. Rotation between IMU and Camera

Figure 5 shows the several frames of reference that need to be considered. The inertial measurements have to be mapped to the camera frame of reference. If the alignment between them is unknown, calibration is required.

Both sensors can be used to measure the vertical direction, so that the rigid transformation between the IMU frame of reference $\{\mathcal{I}\}$ and the camera frame of reference $\{\mathcal{C}\}$ can be determined. When the IMU sensed acceleration is equal in magnitude to gravity, the sensed direction is the vertical. The camera vertical direction can be taken from the vanishing point of either a specific calibration target, such as a chessboard placed vertically, or from some known scene vertical edges. However, camera calibration is required to obtain the correct 3D orientation of the vanishing points.

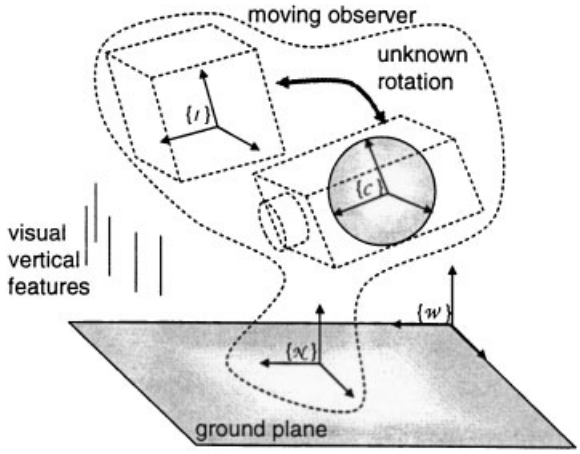


**Figure 5.** Camera $\{\mathcal{C}\}$, IMU $\{\mathcal{I}\}$, mobile system $\{\mathcal{N}\}$, and world fixed $\{\mathcal{W}\}$ frames of reference.

If $n$ observations are made for distinct camera positions, recording the vertical reference provided by the inertial sensors and the vanishing point of scene vertical features, the absolute orientation can be determined using Horn's method.[29] Since we are only observing a 3D direction in space, we can only determine the rotation between the two frames of reference.

Let ${}^{\mathcal{I}}\mathbf{v}_i$ be a measurement of the vertical by the inertial sensors and ${}^{\mathcal{C}}\mathbf{v}_i$ the corresponding measurement made by the camera derived from some scene vanishing point. We want to determine the unit quaternion $\mathring{\mathbf{q}}$ that rotates inertial measurements in the inertial sensor frame of reference $\{\mathcal{I}\}$ to the camera frame of reference $\{\mathcal{C}\}$. In the following equations, when multiplying vectors with quaternions, the corresponding imaginary quaternions are implied. We want to find the unit quaternion $\mathring{\mathbf{q}}$ that maximizes

$$\sum_{i=1}^{n} (\mathring{\mathbf{q}}\,{}^{\mathcal{I}}\mathbf{v}_i\mathring{\mathbf{q}}^{*}) \cdot {}^{\mathcal{C}}\mathbf{v}_i\,. \tag{10}$$

Expressing the quaternion product $\mathring{\mathbf{q}}\mathbf{v}_i$ as a matrix multiplication $\mathbb{V}_i\mathring{\mathbf{q}}$, after some manipulation we get

$$\sum_{i=1}^{n} \mathring{\mathbf{q}}^T \, {}^{\mathcal{I}}\mathbb{V}_i^T \, {}^{\mathcal{C}}\mathbb{V}_i \, \mathring{\mathbf{q}}; \tag{11}$$

factoring out $\mathring{\mathbf{q}}$ we get

$$\mathring{q}^T \left( \sum_{i=1}^{n} {}^{\mathcal{I}}\mathbb{V}_i^T \, {}^{\mathcal{C}}\mathbb{V}_i \right) \mathring{\mathbf{q}}. \tag{12}$$

So we want to find $\mathring{\mathbf{q}}$ such that

$$\max \mathring{\mathbf{q}}^T \mathbf{N} \mathring{q}, \tag{13}$$

where $\mathbf{N} = \Sigma_{i=1}^{n} \, {}^{\mathcal{I}}\mathbb{V}_i^T \, {}^{\mathcal{C}}\mathbb{V}_i$. Matrix $\mathbf{N}$ can be expressed using the sums for all $i$ of all nine pairing products of the components of the two vectors ${}^{\mathcal{I}}\mathbf{v}_i$ and ${}^{\mathcal{C}}\mathbf{v}_i$. The sums contain all the information that is required to find the solution. Since $\mathbf{N}$ is a symmetric matrix, the solution to this problem is the four-vector $\mathbf{q}_{\max}$ corresponding to the largest eigenvalue $\lambda_{\max}$ of $\mathbf{N}$—see ref. 29 for details. Results of this calibration method are presented in ref. 30.

### 4.3. Using the Inertial Vertical Reference

The vertical reference $\hat{\mathbf{n}}$ corresponds to the *north pole* of the unit sphere. A set of world vertical features will project to image lines $\mathbf{n}_i$ with a common vanishing point $\mathbf{m}_{\mathbf{vp}} = \hat{\mathbf{n}}$.

Given a single image vanishing point $\mathbf{v}_p = (x, y)$ of a levelled plane, the horizon line is given by

$$n_x x + n_y y + n_z f = 0, \tag{14}$$

where $f$ is the focal distance and $\hat{\mathbf{n}} = (n_x, n_y, n_z)^{\mathsf{T}}$. Since the vanishing line is determined alone by the orientation of the planar surface, the horizon line is the vanishing line of all levelled planes, parallel to the ground plane.

If a ground plane world point $\mathbf{P}$, given in the camera frame of reference $\{\mathcal{C}\}$, is known, the plane equation can be determined and is given by

$$\hat{\mathbf{n}} . \mathbf{P} + d = 0, \tag{15}$$

where $d$ is the distance from the origin to the ground plane, i.e., the system height. In some applications it can be known or imposed by the physical mount, or determined using stereo as shown below.

When detecting world features, a convenient frame of reference has to be established. A moving robot navigation frame of reference $\{\mathcal{N}\}$ can be considered, aligned by the ground pane as shown in Figure 5. The vertical unit vector $\hat{\mathbf{n}}$ and system height $d$ can be used to define $\{\mathcal{N}\}$; by choosing ${}^{\mathcal{N}}\hat{\mathbf{x}}$ to be coplanar with ${}^{\mathcal{C}}\hat{\mathbf{x}}$ and ${}^{\mathcal{C}}\hat{\mathbf{n}}$ in order to keep the same heading,[28] we have

$${}^{\mathcal{N}}\mathbf{P} = {}^{\mathcal{N}}\mathbf{T}_{\mathcal{C}} . {}^{\mathcal{C}}\mathbf{P}, \tag{16}$$

where

$${}^{\mathcal{N}}\mathbf{T}_{\mathcal{C}} = \begin{bmatrix} \sqrt{1-n_x^2} & \dfrac{-n_x n_y}{\sqrt{1-n_x^2}} & \dfrac{-n_x n_z}{\sqrt{1-n_x^2}} & 0 \\ 0 & \dfrac{n_z}{\sqrt{1-n_x^2}} & \dfrac{-n_y}{\sqrt{1-n_x^2}} & 0 \\ n_x & n_y & n_z & d \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{17}$$

If a heading reference is available, then $\{\mathcal{N}\}$ should not be restricted to having ${}^{\mathcal{N}}\hat{\mathbf{x}}$ coplanar with ${}^{\mathcal{C}}\hat{\mathbf{x}}$ and ${}^{\mathcal{C}}\hat{\mathbf{n}}$, but use the known heading.[28] Using the robot's odometry, the inertial sensors, and landmark matching, conversion to the world fixed frame of reference $\{\mathcal{W}\}$ can be accomplished.

### 4.4. Stereo Depth Map Alignment Using Vertical Reference

Stereo vision systems can use correlation based methods to obtain depth maps. With the current technology, real time systems are commercially available. When the vision system is moving the maps have to be fused into a single world map. Before fusing the depth maps, they must be registered to a common referential. This can be done using data fitting alone, or aided by known parameters or restrictions on the way the measurements were made.

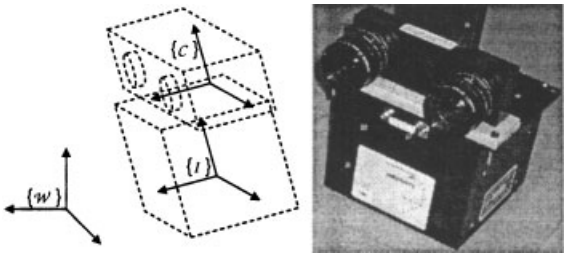


**Figure 6.** Frames of reference and stereo vision system with inertial measurement unit.

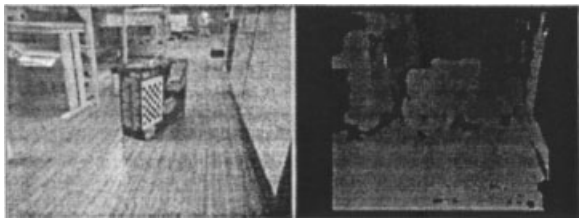

**Figure 7.** Observed scene and depth map obtained with SVS.[31]

In order to obtain depth maps with known vision system pose, the stereo vision system was mounted onto an inertial measurement unit (IMU), as shown in Figure 6. To compute range from stereo images we are using the SRI stereo engine[31] with the small vision system (SVS) software and the MEGA-D digital stereo head, shown in Figure 6.

The depth maps are given in the right camera frame of reference, with the $Z$ axis pointing forward along the optical axis. The depth map is given by a pencil of rays with known depth from the origin (Figure 7).

Using the vertical reference, the depth maps can be segmented to identify horizontal and vertical features. The aim is on having a simple algorithm suitable for a real-time implementation. Since we are able to map the points to an inertial reference frame, planar leveled patches will have the same depth $z$, and vertical features the same $xy$, allowing simple feature segmentation using histogram local peak detection. Figure 8 summarizes the proposed depth map segmentation method.

The depth map points are mapped to the world frame of reference. In order to detect the ground plane, a histogram is performed for the different heights. The histogram's lower local peak, $z_{gnd}$, is used as the reference height for the ground plane. Figure 9 shows some results of ground plane detection and depth map rectification.

## 5. COMBINING DYNAMIC INERTIAL CUES WITH VISION

Inertial sensors can only provide direct measurements of angular velocity $\boldsymbol{\omega}$ and, after subtracting gravity, linear acceleration $\mathbf{a}$. Angular position required to subtract gravity is obtained from integration over time, with unbounded error buildup. Linear velocity and position is again done by integration over time with the associated error accumulation. As previously described, some heuristics can be applied to reset or bound some of the error drift.

From (8) we see how velocities $\mathbf{t}$ and $\boldsymbol{\omega}$ are projected onto the image. Inertial angular velocity measurements, being directly measured, should be used with more confidence than linear velocities.

### 5.1. Image Focus of Expansion

When the camera is moving with linear velocity $\mathbf{t}$ and not rotating, from (8) we see that the image point

$$\mathbf{m}_{\mathrm{FOE}} = \frac{\mathbf{t}}{\|\mathbf{t}\|} \tag{18}$$

will have no motion, i.e., $\dot{\mathbf{m}}_{\mathrm{FOE}} = 0$, and all others will be expanding or contracting to this point. This point

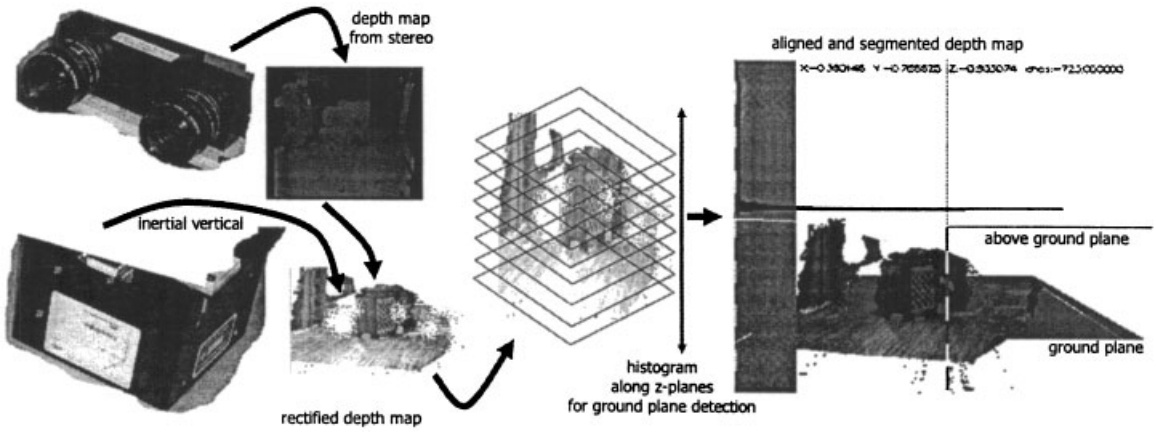


**Figure 8.** Summary of depth map vertical alignment method.

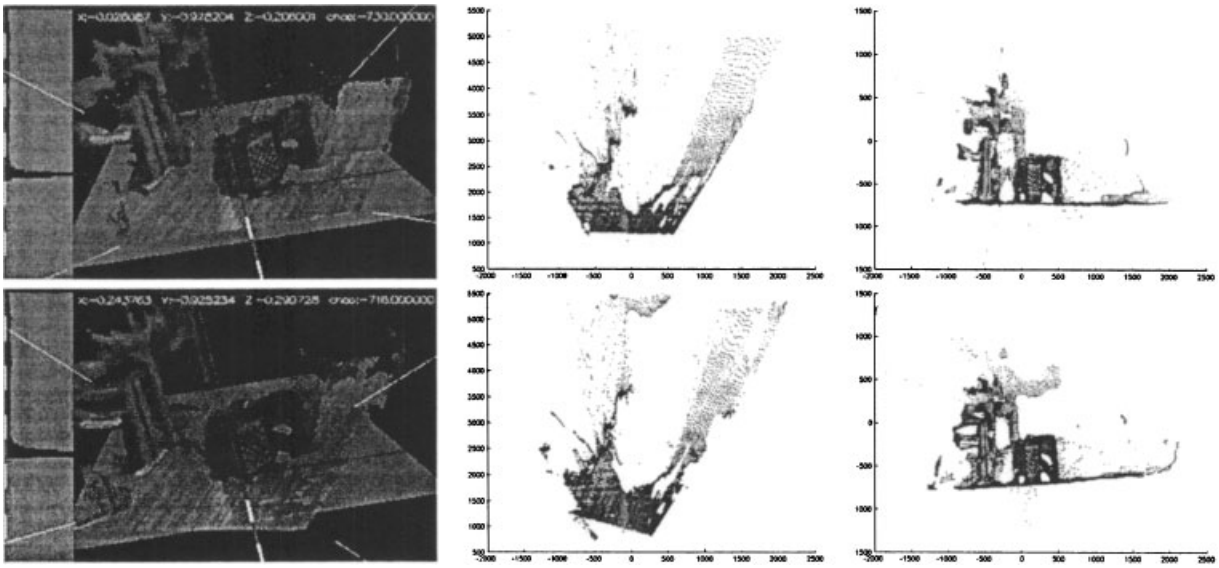

**Figure 9.** On the left the graphical front-end of the implemented system, showing the height histogram for ground plane detection, the detected plane and 3D segmented depth map; on the right the top and front view of the aligned segmented depth maps that only require a translation $(t_x, t_y)$ and rotation $\theta$ to be correctly fused.

is known as the the image focus of expansion (FOE). When the system is also rotating, the FOE will have depth independent velocity

$$\dot{\mathbf{m}}_{\mathrm{FOE}} = -\boldsymbol{\omega} \times \mathbf{m}_{\mathrm{FOE}} = -\boldsymbol{\omega} \times \frac{\mathbf{t}}{\|\mathbf{t}\|}. \tag{19}$$

The FOE can be found using the inertial data alone, provided that the system has been calibrated.

### 5.2. Image Center of Rotation

When the camera is moving with angular velocity $\mathbf{t}$ and no linear translation, from (8) we see that the image point

$$\mathbf{m}_{\mathrm{COR}} = \frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|} \tag{20}$$

will have no motion, i.e., $\dot{\mathbf{m}}_{\mathrm{COR}} = 0$, and all others will be rotating around this point. This point is known as the image center of rotation (COR). When the system is also translating at velocity $\mathbf{t}$, the COR will have depth dependent velocity

$$\dot{\mathbf{m}}_{\mathrm{COR}} = \frac{1}{\|\mathbf{P}_{\mathrm{FOE}}\|}((\mathbf{t}.\mathbf{m}_{\mathrm{COR}})\mathbf{m}_{\mathrm{COR}} - \mathbf{t}), \tag{21}$$

where $\mathbf{P}_{\mathrm{FOE}}$ in the 3D point in view along the image ray given by $\mathbf{m}_{\mathrm{COR}}$. The COR can be easily defined using the inertial data alone, provided that the system has been calibrated using the procedure described in Section 4. The definition of the FOE and COR can be useful during visual based navigation tasks.

### 5.3. Registering Depth Maps

With the vision system moving, the acquired depth maps have to be registered to a common frame of reference. After the alignment using the vertical reference and subsequent ground plane detection, the registration is a 2D problem; only a translation $(t_x, t_y)$ and rotation $\theta$ are needed (see Figure 9).

An approximation to these 2D parameters can be found by projecting the inertial sensed parameters onto the level plane. These allow registering dynamic depth maps, with moving objects, to a common frame of reference.

## 6. CONCLUSIONS

This paper has presented a framework for the combination of inertial and visual sensing modalities.

Keeping track of the vertical direction provides a valuable spatial reference. Results were shown of stereo depth map alignment using the vertical reference. The depth map points are mapped to a vertically

aligned world frame of reference. In order to detect the ground plane, a histogram is performed for the different heights. Taking the ground plane as a reference plane for the acquired maps, the fusion of multiple maps reduces to a 2D translation and rotation problem. The dynamic inertial cues can be used as a first approximation for this transformation, allowing a fast depth map registration method.

The definition of the FOE and COR can be done from the inertial cues, and used during visual based navigation tasks.

Future work will address the fusion of optical flow computation with the inertial ego-motion estimate. The image optical flow imposes a further restriction to bound the drift in the inertial estimated ego-motion. The computed depth from flow and known ego-motion can be combined with the stereo correlation computed depth, producing a more robust 3D reconstruction technique.

## REFERENCES


1. H. Carpenter, Movements of the eyes, London Pion Limited, London, 1988, 2nd ed.
2. A. Berthoz, The brain's sense of movement, Harvard UP, Harvard, 2000.
3. J. Lobo and J. Dias, Vision and inertial sensor cooperation, using gravity as a vertical reference, IEEE Trans Pattern Analy Mach Intell 25:(12) (2003), in press.
4. J. Lobo, C. Queiroz, and J. Dias, World feature detection and mapping using stereovision and inertial sensors, Robot Auton Syst Elsevier Science, 44:(1) (2003), 69–81.
5. R.P.G. Collinson, Introduction to avionics, Chapman & Hall, New York, 1996.
6. G.R. Pitman, Inertial guidance, Wiley, New York, 1962.
7. J.J. Allen, R.D. Kinney, J. Sarsfield, M.R. Daily, J.R. Ellis, J.H. Smith, S. Montague, R.T. Howe, B.E. Boser, R. Horowitz, A.P. Pisano, M.A. Lemkin, W.A. Clark, and T. Juneau, Integrated micro-electro-mechanical sensor development for inertial applications, in Proc 1998 Position Location and Navigation Symposium, April 1998.
8. T. Viéville and O.D. Faugeras, Computation of inertial information on a robot, in Symposium on Robotics Research, edited by H. Miura and S. Arimoto, editors, Fifth Inter MIT, Cambridge, 1989, pp. 57–65.
9. T. Viéville and O.D. Faugeras, Cooperation of the inertial and visual systems, in Traditional and non traditional robotic sensors, vol. F 63 of NATO ASI, edited by T.C. Henderson, Springer-Verlag, Berlin, 1990, pp. 339–350.
10. T. Viéville, F. Romann, B. Hotz, H. Mathieu, M. Buffa, L. Robert, P.E.D.S. Facao, O. Faugeras, and J.T. Audren, Autonomous navigation of a mobile robot using inertial and visual cues, in Intelligent robots and systems, edited by M. Kikode, T. Sato, and K. Tatsuno, Yokohama, 1993.
11. T. Viéville, E. Clergue, and P.E.D. Facao, Computation of ego-motion and structure from visual and inertial sensor using the vertical cue, in *ICCV93*, 1993, pp. 591–598.
12. T. Viéville, A few steps towards 3D active vision, Springer-Verlag, New York, 1997.
13. B. Bhanu, B. Roberts, and J. Ming, Inertial navigation sensor integrated motion analysis for obstacle detection, in Proc 1990 IEEE Int Conf on Robotics and Automation, Cincinnati, OH, 1990, pp. 954–959.
14. F. Panerai and G. Sandini, Visual and inertial integration for Gaze stabilization, in Proc SIRS'97, Stockholm, 1997.
15. F. Panerai and G. Sandini, Oculo-motor stabilization reflexes: integration of inertial and visual information, Neural Networks, 11:(7-8) (1998), 1191–1204.
16. F. Panerai, G. Metta, and G. Sandini, Visuo-inertial stabilization in space-variant binocular systems, Roboti Auton Syste, 30:(1-2) (2000), 195–214.
17. T. Mukai and N. Ohnishi, The recovery of object shape and camera motion using a sensing system with a video camera and a gyro sensor, in Proc Seventh Int Conf on Computer Vision (ICCV'99), Kerkyra, Greece, September 1999, pp. 411–417.
18. T. Mukai and N. Ohnishi, Object shape and camera motion recovery using sensor fusion of a video camera and a gyro sensor, Inf Fusion, 1:(1) (2000), 15–53.
19. R. Kurazume and S. Hirose, Development of image stabilization system for remote operation of walking robots, in Proc 2000 IEEE Int Conf on Robotics and Automation, San Francisco, CA, April 2000, pp. 1856–1860.
20. S.R. Coorg, Pose imagery and automated three-dimensional modeling of urban environments, PhD thesis, Massachusetts Institute of Technology, September 1998.
21. S. You, U. Neumann, and R. Azuma, Hybrid inertial and vision tracking for augmented reality registration, in Proc IEEE Virtual Reality '99, Houston, Texas, March 1999, pp. 260–267.
22. W.A. Hoff, K. Nguyen, and T. Lyon, Computer vision-based registration techniques for augmented reality, in Proc of Intelligent Robots and Computer Vision, November 1996, pp. 538–548.
23. L.-Chai, W.A. Hoff, and T. Vincent, 3-D motion and structure estimation using inertial sensors and computer vision for augmented reality, Presence: Teleop Virt Environ 11:(5) (2002), 474–492.
24. E.D. Dickmanns, Vehicles capable of dynamic vision: a new breed of technical beings? Artif Intelli 103 (1998), 49–76.
25. K. Kanatani, Geometric computation for machine vision, Oxford UP, Oxford, 1993.
26. J. Stolfi, Oriented projective geometry, a framework for geometric computations, Boston Academic, Boston, 1991.
27. K.K. Gillingham and F.H. Previc, Spatial orientation in flight, 2nd ed., chapter 11, Williams and Wilkins, Baltimore, 1996.
28. J. Lobo, Inertial sensor data integration in computer vision systems, Master's thesis, University of Coimbra, April 2002.

29. B.K.P. Horn, Closed-form solution of absolute orientation using unit quaternions, J Opt Soc Am 4:(4) (1987), 629–462.
30. J. Alves, J. Lobo, and J. Dias, Camera-inertial sensor modelling and alignment for visual navigation, in Proc 11th Int Conf on Advanced Robotics, Coimbra, Portugal, July 2003, pp. 1693–1698.
31. K. Konolige, Small vision systems: hardware and implementation, in 8th Int Symposium on Robotics Research, Hayama, Japan, October 1997.