Vision and Inertial Sensor Cooperation Using Gravity as a Vertical Reference

Jorge Lobo and Jorge Dias

Abstract—This paper explores the combination of inertial sensor data with vision. Visual and inertial sensing are two sensory modalities that can be explored to give robust solutions on image segmentation and recovery of 3D structure from images, increasing the capabilities of autonomous robots and enlarging the application potential of vision systems. In biological systems, the information provided by the vestibular system is fused at a very early processing stage with vision, playing a key role on the execution of visual movements such as gaze holding and tracking, and the visual cues aid the spatial orientation and body equilibrium. In this paper, we set a framework for using inertial sensor data in vision systems, and describe some results obtained. The unit sphere projection camera model is used, providing a simple model for inertial data integration. Using the vertical reference provided by the inertial sensors, the image horizon line can be determined. Using just one vanishing point and the vertical, we can recover the camera's focal distance and provide an external bearing for the system's navigation frame of reference. Knowing the geometry of a stereo rig and its pose from the inertial sensors, the collineation of level planes can be recovered, providing enough restrictions to segment and reconstruct vertical features and leveled planar patches.

Index Terms—Image processing and computer vision, edge and feature detection, sensor fusion.

1 Introduction

In humans and in animals, the vestibular system in the inner ear gives inertial information essential for navigation, orientation, body posture control, and equilibrium. In humans, this sensorial system is crucial for several visual tasks and head stabilization. It is well-known that the information provided by the vestibular system is used during the execution of visual movements such as gaze holding and tracking, as described by Carpenter [1]. Neural interactions of human vision and the vestibular system occur at a very early processing stage [2], [3]. The inertial information enhances the performance of the vision system and the visual cues aid the spatial orientation and body equilibrium.

Inertial sensors explore intrinsic properties of body motion. From the principle of generalized relativity of Einstein, we know that only the specific force on one point and the angular instantaneous velocity, but no other quantity concerning motion and orientation with respect to the rest of the universe can be measured from physical experiments inside an isolated closed system. Therefore, from inertial measurements, one can only determine an estimate for linear acceleration and angular velocity. Linear velocity and position and angular position can be obtained by integration. Inertial navigation systems (INS) implement this process of obtaining velocity and position information from inertial sensor measurements.

Internal sensing using inertial sensors is very useful in mobile robotic systems since it is not dependent on any external references, except for the gravity field which does provide an external reference. Artificial vision systems can provide better perception of the robot's environment by using the inertial sensors 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.

Micromachining enabled the development of low-cost single chip inertial sensors. These can be easily incorporated alongside the camera's imaging sensor, providing an artificial vestibular system. The noise level of these sensors is not suitable for inertial navigation systems, but their performance is similar to biological inertial sensors and can play a key role in artificial vision systems. In our work, we explore some aspects of inertial and vision sensing integration. Fig. 1 shows some of the data available from the two sensors and processing systems, setting a framework for possible combination of inertial and vision sensing.

The 3D structured world is observed by the visual sensor and its pose and motion parameters are directly measured by the inertial sensors. These motion parameters can also be inferred from the image flow and known scene features [4]. Combining the two sensing modalities simplifies the 3D reconstruction of the observed world. The inertial sensors also provide important cues about the observed scene structure, such as vertical and horizontal references.

1.1 Related Work

Inertial sensors have typically been used for navigation in aerospace and naval applications [5], [6]. The electronic and silicon micromachining development, pushed by the needs of the automotive industry, brought about low cost, batch fabricated silicon sensors. New micromachined sensors are being developed, aiming at having a single chip inertial system to be integrated in inertial aided GPS navigation

- *The authors are with the ISR-Institute of Systems and Robotics and the Electrical and Computer Engineering Department of the University of Coimbra, 3030-290 Coimbra, Portugal. E-mail: {jlobo, jorge}@isr.uc.pt.*

Manuscript received 25 July 2002; revised 14 Feb. 2003; accepted 8 July 2003.
Recommended for acceptance by R. Sharma.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 117015.

0162-8828/03/$17.00 © 2003 IEEE Published by the IEEE Computer Society

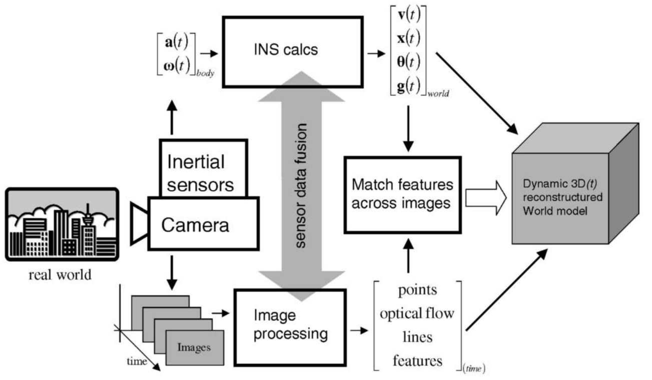

$\begin{bmatrix}\mathbf{a}(t)\\ \boldsymbol{\omega}(t)\end{bmatrix}_{body}$
INS calcs
$\begin{bmatrix}\mathbf{v}(t)\\ \mathbf{x}(t)\\ \boldsymbol{\theta}(t)\\ \mathbf{g}(t)\end{bmatrix}_{world}$
Inertial sensors
Camera
real world
sensor data fusion
Match features across images
Dynamic 3D(t) reconstructured World model
Images
time
Image processing
points
optical flow
lines
features
(time)

Fig. 1. Combining inertial and vision sensing.

systems [7]. This development has enabled many new applications for inertial sensors, namely in robotics and computer vision.

One application that has matured in consumer products is camera vibration compensation. Video camera image stabilization can be done using camera motion detection followed by image correction [8].

Viéville and Faugeras proposed the use of an inertial system based on low cost sensors for mobile robots [9], and studied the cooperation of the inertial and visual systems in mobile robot navigation by using the vertical cue taken from the inertial sensors [10], [11], [12]. In Viéville's book [13], he describes the work done on the integration of inertial sensor data in vision systems to rectify images and improve self-motion estimation for 3D structure reconstruction.

An inertial sensor integrated optical flow technique was proposed by Bhanu et al. [14]. Inertial sensors were used to improve optical flow for obstacle detection. The inertial system was used to compensate ego motion of the vehicle, improving interest point selection, matching of the interest points, and the subsequent motion detection, tracking, and obstacle detection.

Panerai and Sandini used a low cost gyroscope for gaze stabilization of a rotating camera, and compared the camera rotation estimate given by image optical flow with the gyro output [15], [16]. They also studied the integration of inertial and visual information in binocular vision systems [17].

Mukai and Ohnishi studied the recovery of 3D shape from an image sequence using a video camera and a gyro sensor [18], [19]. Rotation and translation have similar effects on the image, leading to unreliable recovery. The gyro output is used to discriminate both situations and improve the accuracy of the 3D shape recovery.

More recently, Kurazume and Hirose have used inertial sensors for image stabilization of remote legged robots and attitude estimation [20].

A strong application for inertial aided vision systems is virtual reality modeling and augmented reality. By using *pose imagery* (i.e., images with known orientation and position obtained by inertial sensors and GPS), Coorg and his colleagues at the MIT Media Lab [21] use mosaicing and other techniques to perform an automated three-dimensional modeling of urban environments. You used a hybrid inertial and vision tracking algorithm for augmented reality registration [22]. Inertial sensors and cameras were used on a head mount system providing 3D motion and structure estimation for augmented reality [23].

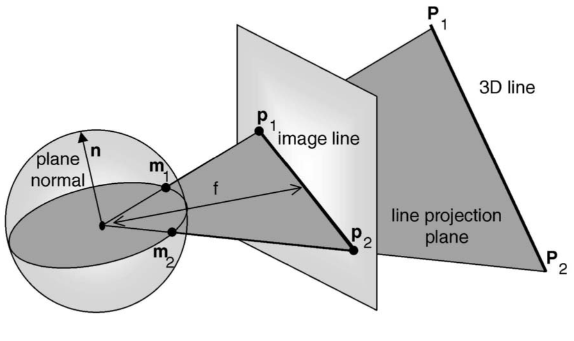


Fig. 2. Line projection onto unit sphere.

Dickmanns has also incorporated inertial sensors in his vision system for automated vehicles [24]. The system uses feedback from the estimated state to guide the vision feature trackers. The inertial sensor-based ego state estimation has negligible time delays and includes perturbations which must be taken into account by the vision system.

1.2 Our Work

In this paper, we present some results obtained using inertial sensor data in vision systems. The unit sphere projection camera model is used, providing a simple model for inertial data integration. Using the vertical reference provided by the inertial sensors, the image horizon line can be determined. Using just one vanishing point, we can recover the camera's focal distance. In a typical indoor corridor scene, the vanishing point can also provide an external bearing for the system's navigation frame of reference. Knowing the geometry of a stereo rig and its pose from the inertial sensors, the collineation of level planes can be recovered, providing enough restrictions to segment and reconstruct vertical features and leveled planar patches. Parts of this work have been previously published in conference proceedings [25], [26], [27], [28], [29], [30].

2 Vision Background and Geometric Framework

2.1 Projection onto Unit Sphere

The pinhole camera model derives from the camera's geometry and considers the projection of world points onto a plane, but the projection need not be onto a plane. Consider a unit sphere around the optical center with the images 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 Fig. 2. The image plane touches the sphere at the equator, and this point defines, on the image plane, the image center. Using the unit sphere gives a more general model for central perspective and provides an intuitive visualization of projective geometry [31]. It also has numerical advantages when considering points at infinity, such as vanishing points.

As shown in Fig. 2, a world point $\boldsymbol{P}$ will project on the image plane as $\boldsymbol{p}$ and can be represented by the unit vector $\boldsymbol{m}$ placed at the sphere's center, the optical center of the camera. With image centered coordinates $\boldsymbol{p} = (u, v)^{\top}$, we have

$$\boldsymbol{P} \rightarrow \boldsymbol{m} = \frac{\boldsymbol{P}}{||\boldsymbol{P}||} = \frac{1}{\sqrt{u^2 + v^2 + f^2}} \begin{bmatrix} u \\ v \\ f \end{bmatrix}. \tag{1}$$

Note that $\boldsymbol{m} = (m_1, m_2, m_3)^{\top}$ is a unit vector and the projection is not defined for $\boldsymbol{P} = (0, 0, 0)^{\top}$. Projection onto the unit sphere is related to projection onto a plane of image point $\boldsymbol{p} = (u, v)^{\top}$ by

$$(u, v)^{\top} = \left(f\frac{m_1}{m_3}, f\frac{m_2}{m_3} \right)^{\top}. \tag{2}$$

Given f, the projection onto a sphere can be computed from the projection onto a plane and conversely. To avoid ambiguity, m_3 is forced to be positive, so that only points on the image side hemisphere are considered.

Image lines can also be represented in a similar way. Any image line defines a plane with the center of projection, as shown in Fig. 2. A vector $\boldsymbol{n}$ normal to this plane uniquely defines the image line and can be used to represent the line.

For a given image line $au + bv + c = 0$, the unit vector is given by

$$\boldsymbol{n} = \frac{1}{\sqrt{a^2 + b^2 + (c/f)^2}} \begin{bmatrix} a \\ b \\ c/f \end{bmatrix}. \tag{3}$$

As seen in Fig. 2, we can write the unit vector of an image line with points $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ as

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

Image points $\boldsymbol{m}$ and $\boldsymbol{m}'$ are said to be conjugate to each other if $\boldsymbol{m}.\boldsymbol{m}' = 0$. In image coordinates, we have that image points $(u, v)^{\top}$ and $(u', v')^{\top}$ are conjugate to each other if

$$uu' + vv' + f^2 = 0, \tag{5}$$

and the projective lines passing through each point and the center of projection are orthogonal.

2.2 Vanishing Points and Vanishing Lines

Since the perspective projection maps a 3D world onto a plane or planar surface, phenomena that only occur *at infinity* will project to very finite locations in the image. Parallel lines only meet at infinity, but as seen in Fig. 3, 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 a unit vector $\boldsymbol{m}$ has, when projected, a *vanishing point* with unit sphere vector $\pm\boldsymbol{m}$. Since the vanishing point is only determined by the 3D orientation of the space line, projections of parallel space lines intersect at a common vanishing point.

A planar surface with a unit normal vector $\boldsymbol{n}$, not parallel to the image plane has, when projected, a *vanishing line* with unit sphere vector $\pm\boldsymbol{n}$. Since the vanishing line is determined alone by the orientation of the planar surface, then the projections of planar surfaces parallel in the scene define a common vanishing line. A vanishing line is a set of all vanishing points corresponding to the lines that belong to the set of parallel planes defining the vanishing line.

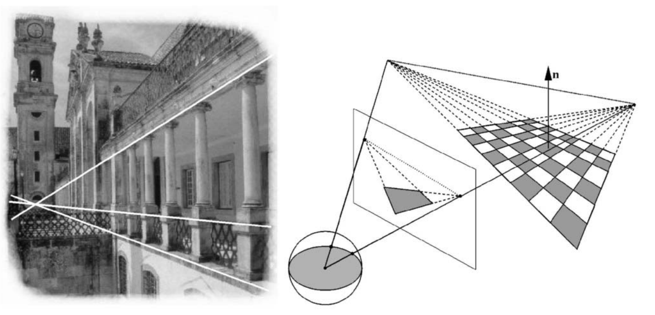


Fig. 3. Picture of *Via Latina* at Coimbra University showing a vanishing point and vanishing line of a planar surface.

In an image, the horizon line can be found by having two distinct vanishing points as seen in Fig. 3. With a suitable calibration target (e.g., a leveled square with well defined edges), the horizon line can be determined.

If the vanishing points, $(u, v)^{\top}$ and $(u', v')^{\top}$, correspond to orthogonal sets of parallel lines, they are conjugate to each other and, from (5), we have

$$f = \sqrt{-uu' - vv'}. \tag{6}$$

Therefore, with two vanishing points corresponding to orthogonal sets of parallel lines, focal distance f can be determined [31].

3 Data from Inertial Sensors

3.1 Inertial Navigation Principles

At the most basic level, an inertial system simply performs a double integration of sensed acceleration over time to estimate position. Assuming a set of accelerometers measuring acceleration along three orthogonal axis, we have

$$\boldsymbol{\Delta}_{position} = \boldsymbol{x} = \int \dot{\boldsymbol{x}}\, dt = \iint \ddot{\boldsymbol{x}}\, dt = \iint \boldsymbol{a}_{sensed} dt, \tag{7}$$

where $\boldsymbol{x}$ is the position, $\dot{\boldsymbol{x}}$ the velocity, and $\ddot{\boldsymbol{x}}$ the acceleration vectors.

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 gimballed systems, the accelerometers are kept in alignment with the navigation frame of reference, using the gyros to control a stabilized platform. In strap-down systems, the gyros measure the body rotation rate and the sensed accelerations are computationally converted to the navigation frame of reference. Before integration, gravity $\boldsymbol{g}$ must be subtracted from the sensed acceleration. Fig. 4 shows a block diagram of a strapdown inertial navigation system. The Inertial Measurement Unit (IMU) has three acceleromters and three gyrometers.

The mechanization of this rigid body angular motion has to account for the noncommutativity of finite rotations, mathematical singularities, and numerical instability. Shuster [32] discusses the various derivations for the rotation

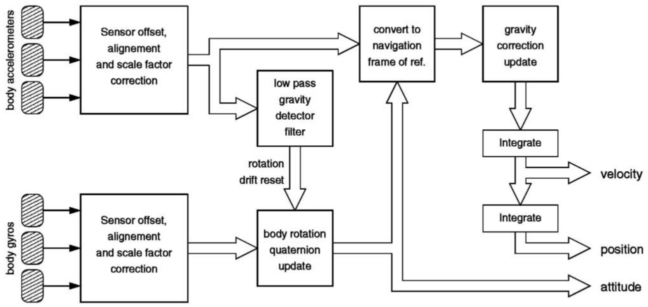

body accelerometers
Sensor offset, alignement and scale factor correction
low pass gravity detector filter
convert to navigation frame of ref.
gravity correction update
Integrate
velocity
Integrate
position
attitude
rotation drift reset
body gyros
Sensor offset, alignement and scale factor correction
body rotation quaternion update

Fig. 4. Simplified strap-down inertial navigation system.

vector, and a complete mechanization using quaternions is presented by Savage [33]. Table 1 summarizes the data that can be obtained from the inertial sensors.

Recent development in micromechanical devices has lead to some new low-cost accelerometers and gyroscopes. Strap-down systems based on these low-cost inertial sensors offer low performance, namely in accumulated drift over time, making them unsuitable for high performance inertial systems, but can still be useful in some mobile robotic applications. The inertial system can be used to provide short-term accurate relative positioning and, combined with some other external absolute positioning system to bound the INS drift error, provide a suitable navigation system.

To cope with the accumulated drift, some assumptions can be made on the system's dynamics. If the norm of the sensed acceleration is about $9.8\ m.s^{-2}$, then we can assume that the accelerometers only measure gravity $\boldsymbol{g}$ and the attitude can be directly determined, resetting the accumulated drift in the attitude computation. Assuming pure rotations never occur, we could also adjust the gyro offset since they tend do drift with time and temperature. A low threshold can also be applied to the system, assuming that the robot never accelerates or rotates below a certain value, preventing the error accumulation in the rotation update and position integration.

3.2 Performance of Human Inertial Sensors

It is important to have some idea of the performance of the human inertial sensors to better evaluate the suitability of inertial sensors in some robotic applications. But, 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 [3]. Reasonable threshold values for perception of angular acceleration are 0.14, 0.5, and $0.5\ deg.s^{-2}$ for yaw, roll, and pitch motions, respectively. A $1.5\ deg$ change in direction of applied gravity force is perceptible by the otolith organs under ideal conditions. 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 [30]. The inertial system prototype built for this work, using low cost sensors, has gyros with $0.1\ deg.s^{-1}$ resolution, and accelerometers with $0.005\ g$ resolution. Notice that the gyros measure angular velocity and not angular acceleration.

TABLE 1
Data from Inertial Sensors

differentiation	$\boldsymbol{\varphi} = \ddot{\boldsymbol{\theta}}$ $\boldsymbol{j} = \dot{\boldsymbol{a}}$	angular acceleration rate of linear acceleration (jerk)
direct measurement	$\boldsymbol{\omega} = \dot{\boldsymbol{\theta}}$ $\boldsymbol{a} + \boldsymbol{g} = \ddot{\boldsymbol{x}} + \boldsymbol{g}$	angular velocity linear acceleration + gravity
integration	$\boldsymbol{\theta}$ $\boldsymbol{v} = \dot{\boldsymbol{x}}$	angular position (attitude) linear velocity
double integration	$\boldsymbol{x}$	position

These performances are not suitable for stand alone 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 Inertial Cues with Monocular Vision

4.1 Unit Sphere Vertical Reference from Gravity

The accelerometer data can be used to determine the vision system's attitude. When the system is motionless or subject to constant speed, the accelerometers give the direction of the gravity vector $\boldsymbol{g}$ relative to the camera system frame of reference $\{\mathcal{C}\}$. We can therefore determine the vertical unit vector normal to local ground leveled plane, but rotations within the horizontal plane are not sensed. The measurements $\boldsymbol{a}$ taken by the inertial unit's accelerometers include the sensed gravity vector $\boldsymbol{g}$ summed with the body's acceleration $\boldsymbol{a}_b$:

$$\boldsymbol{a} = -\boldsymbol{g} + \boldsymbol{a}_b. \tag{8}$$

Notice that the accelerometer will measure the reactive (upward) force to gravity. Assuming the system is motionless, then $\boldsymbol{a}_b = 0$ and the measured acceleration $\boldsymbol{a}$ gives the gravity vector in the system's frame of reference. So, with a_x, a_y, and a_z being the accelerometer measurements along each axis, the vertical unit vector will be given by

$$\hat{\boldsymbol{n}} = \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = -\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|} = \frac{1}{\sqrt{a_x^2 + a_y^2 + a_z^2}} \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix}. \tag{9}$$

As explained in the previous section, by performing the rotation update using the IMU gyro data, gravity can be separated from the sensed acceleration. In this case, $\hat{\boldsymbol{n}}$ 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 frame of reference and has to be converted to the camera frame of reference. Only the rotation is relevant, for a single perfectly aligned camera it is null; otherwise, it can be calibrated as described below. In the case of having a stereo camera rig, either the verge of each camera and IMU relative pose is also known, or some calibration has to be performed.

This vertical reference will be used together with image cues to calibrate the vision sensor and detect world features such as vertical lines and the ground plane, as presented in the following sections.

4.2 Vertical Reference Error Analysis

Considering a linear model for the accelerometers, we have

$$\boldsymbol{a}_{measured} = \boldsymbol{M}\boldsymbol{a}_{real} + \boldsymbol{b}, \tag{10}$$

where M incorporates scale factor, cross axis sensitivity inherent to the sensing element and also due to sensor misalignment, and $\boldsymbol{b}$ offset bias. Estimates for M and $\boldsymbol{b}$ might be provided by the manufacturer or can be obtained by sensor calibration.

But, temperature drift, power supply ripple interference, and thermal noise will always degrade the signal, and M and $\boldsymbol{b}$ will not remain static. Part of this noise can be taken as having zero mean, but, for instance, temperature drift does not, and temperature compensation might be required in some applications.

Since we are measuring gravity, any mechanical vibrations and oscillations will introduce additional error. Low pass filtering has to be used and, for more dynamic situations, gyro rotation update is required to have stabilized gravity direction. In some applications, using magnetic sensors with accelerometers provides a good solution for pose tracking, exploring the different dynamics and noise characteristics of each sensor [34].

The inertial system prototype built at our lab [25] (Fig. 7) uses a signal conditioned three-axis accelerometer [35], [36]. Bias and cross-axis sensitivity calibration data was available and used.

A set of 6,400 measurements were taken with the sensor at rest with no filtering. The obtained covariance matrix was

$$V(\boldsymbol{n}) = \begin{bmatrix} 0.5873 & -0.0069 & 0.0102 \\ -0.0069 & 0.5675 & 0.0071 \\ 0.0102 & 0.0071 & 0.0003 \end{bmatrix} \times 10^{-4}, \tag{11}$$

with eigenvalues

$$\sigma_1^2 = 0.5895 \times 10^{-4} \geq \sigma_2^2 = 0.5655 \times 10^{-4} \geq 0. \tag{12}$$

The root-mean-square angle error is given by

$$\sigma_\theta = tan^{-1}\left(\sqrt{trV(\boldsymbol{n})}\right) = tan^{-1}\left(\sqrt{\sigma_1^2 + \sigma_2^2}\right) \tag{13}$$

and, for this data set, $\sigma_\theta = 0.6130\ deg$. Low pass filtering the accelerometer data improved the estimate significantly, lowering the error to $\sigma_\theta = 0.1907\ deg$ when applying a Butterworth fifth order filter with $10\ Hz$ cutoff frequency.

Another set of measurements was made on a mobile robot performing back-and-forth motion with no filtering. The obtained covariance matrix was

$$V(\boldsymbol{n}) = \begin{bmatrix} 0.6928 & 0.1094 & 0.3086 \\ 0.1094 & 0.7763 & 0.0183 \\ 0.3086 & 0.0183 & 0.1388 \end{bmatrix} \times 10^{-4}, \tag{14}$$

with eigenvalues

$$\sigma_1^2 = 0.9144 \times 10^{-4} \geq \sigma_2^2 = 0.6934 \times 10^{-4} \geq \sigma_3^2 = 0.10 \times 10^{-7} \tag{15}$$

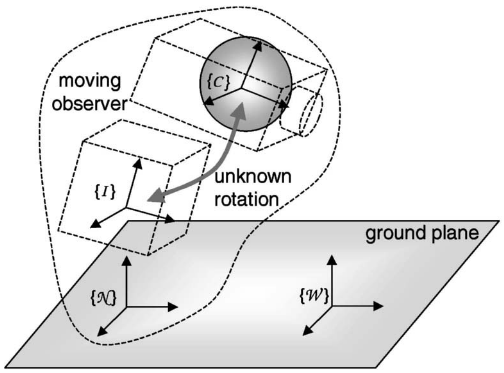


Fig. 5. Camera $\{\mathcal{C}\}$, IMU $\{\mathcal{I}\}$, mobile system $\{\mathcal{N}\}$, and world fixed $\{\mathcal{W}\}$ frames of reference.

and, for this data set, the expected angle error $\sigma_\theta = 0.7265\ deg$. Low pass filtering the accelerometer data improved the estimate significantly, lowering the error to $\sigma_\theta = 0.4611\ deg$.

4.3 Rotation between IMU and Camera

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

In order to determine the rigid transformation between the IMU frame of reference $\{\mathcal{I}\}$ and the camera frame of reference $\{\mathcal{C}\}$, both sensors are used to measure the vertical direction. When the IMU sensed acceleration is equal in magnitude to gravity, the sensed direction is the vertical. For the camera, either using a specific calibration target, such as a chessboard placed vertically, or assuming the scene has enough predominant vertical edges, the vertical direction can be taken from the corresponding vanishing point. However, camera calibration is needed to obtain the correct 3D orientation of the vanishing points.

If n observations are made for distinct system 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 [37]. Since we are only observing a 3D direction in space, we can only determine the rotation between the two frames of reference. Results are given in [38].

4.4 Vanishing Point of Vertical Lines

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

4.5 Horizon Line

The horizon line can be found by having two distinct vanishing points of a leveled plane. Knowing the vertical in the camera's referential and the focal distance, an artificial horizon line also can also be traced with a single vanishing point. A planar surface with a unit normal vector $\hat{\boldsymbol{n}}$, not parallel to the image plane has, when projected, a *vanishing line* given by

$$n_x u + n_y v + n_z f = 0, \tag{16}$$

where f is the focal distance, u and v image coordinates, and $\hat{\boldsymbol{n}} = \left(n_x, n_y, n_z\right)^\top$. Since the vanishing line is determined alone by the orientation of the planar surface, the horizon

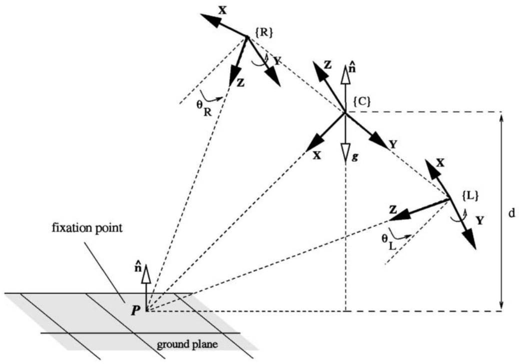


Fig. 6. Ground plane point P_f fixated by stereo system.

line is the vanishing line of all leveled planes, parallel to the ground plane.

4.6 Ground Plane

Consider a world point ${}^{\mathcal{C}}\boldsymbol{P}$, given in a camera centered referential $\{\mathcal{C}\}$, that belongs to the ground plane. The plane equation is given by

$$ {}^{\mathcal{C}}\hat{\boldsymbol{n}}.{}^{\mathcal{C}}\boldsymbol{P} + d = 0, \tag{17}$$

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. The ground plane can therefore be determined in the camera system frame of reference $\{\mathcal{C}\}$.

4.7 Robot Navigation Frame of Reference

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 plane as shown in Fig. 5. The vertical unit vector $\hat{\boldsymbol{n}}$ and system height d can be used to define $\{\mathcal{N}\}$, by choosing ${}^{\mathcal{N}}\hat{\boldsymbol{x}}$ to be coplanar with ${}^{\mathcal{C}}\hat{\boldsymbol{x}}$ and ${}^{\mathcal{C}}\hat{\boldsymbol{n}}$ in order to keep the same heading, we have

$$ {}^{\mathcal{N}}\boldsymbol{P} = {}^{\mathcal{N}}\boldsymbol{T}_{\mathcal{C}}.{}^{\mathcal{C}}\boldsymbol{P}, \tag{18}$$

where

$$ {}^{\mathcal{N}}\boldsymbol{T}_{\mathcal{C}} = \begin{bmatrix} \sqrt{1-n_x^2} & \frac{-n_x n_y}{\sqrt{1-n_x^2}} & \frac{-n_x n_z}{\sqrt{1-n_x^2}} & 0 \\ 0 & \frac{n_z}{\sqrt{1-n_x^2}} & \frac{-n_y}{\sqrt{1-n_x^2}} & 0 \\ n_x & n_y & n_z & d \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{19}$$

If a heading reference is available, then $\{\mathcal{N}\}$ should not be restricted to having ${}^{\mathcal{N}}\hat{\boldsymbol{x}}$ coplanar with ${}^{\mathcal{C}}\hat{\boldsymbol{x}}$ and ${}^{\mathcal{C}}\hat{\boldsymbol{n}}$, but use the known heading reference [30]. As previously seen, vanishing points $\hat{\boldsymbol{m}}_i$ of leveled planes are orthogonal to the vertical $\hat{\boldsymbol{n}}$, i.e., $\hat{\boldsymbol{m}}_i.\hat{\boldsymbol{n}} = 0$. In scenes of man made environments, the vanishing points can provide a heading reference. Using vanishing point $\hat{\boldsymbol{m}} = \left(m_x, m_y, m_z\right)^\top$ as a reference, we get

$$ {}^{\mathcal{C}}\boldsymbol{T}_{\mathcal{N}} = \begin{bmatrix} m_x & n_y m_z - n_z m_y & n_x & -n_x d \\ m_y & n_z m_x - n_x m_z & n_y & -n_y d \\ m_z & n_x m_y - n_y m_x & n_z & -n_z d \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{20}$$

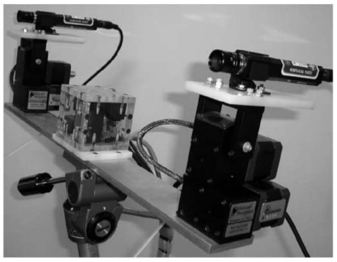


Fig. 7. Stereo camera rig with IMU based on low cost sensors.

Providing suitable vanishing points can be extracted from the scene, we are able to have $\{\mathcal{N}\}$ coherent with the inertial vertical and the scene heading. Using the robot's odometry, the inertial sensors, and landmark matching, conversion to the world fixed frame of reference $\{\mathcal{W}\}$ can be accomplished.

5 Combining Inertial Cues with Stereo Vision

5.1 Ground Plane

As seen above, the vertical reference provides the orientation of the ground plane relative to the camera system. With stereo vision, visual fixation of a ground plane point can be used to determine the ground plane distance [39], [40].

For this stereo system, the camera frame of reference $\{\mathcal{C}\}$ is at the middle of the baseline with x pointing forward, as seen in Fig. 6. Assuming a vision system with controlled symmetric verge angle θ and baseline b, fixated in a point ${}^{\mathcal{C}}\boldsymbol{P}_f$ that belongs to the ground plane, the distance d is given by the projection of ${}^{\mathcal{C}}\boldsymbol{P}_f$ on the gravity vector direction

$$ d = -{}^{\mathcal{C}}\hat{\boldsymbol{n}}.{}^{\mathcal{C}}\boldsymbol{P}_f = -{}^{\mathcal{C}}\hat{\boldsymbol{n}}. \begin{bmatrix} \frac{b}{2}\cot\theta \\ 0 \\ 0 \end{bmatrix} = -n_x \frac{b}{2} \cot\theta, \tag{21}$$

as can easily be seen in Fig. 6. In this figure, there is no lateral inclination, but (21) is valid for any angle since the attitude is given by ${}^{\mathcal{C}}\hat{\boldsymbol{n}}$.

The ground plane can therefore be determined in the camera system frame of reference $\{\mathcal{C}\}$, using the plane orientation, given by the inertial sensors, and the plane height from some a priori knowledge, or by fixating the vision system on a ground plane point. All ground plane geometric parameters are therefore determined. The leveled navigation frame of reference can de shifted to have $Z = 0$ for the ground plane.

5.2 Collineation of Ground Plane Points

To analyze how ground plane points are projected onto the image plane, consider a world point $\boldsymbol{P} = (X, Y, 0, 1)^\top$ that belongs to the ground plane (i.e., $Z = 0$). The projection onto the camera image plane is given by

$$ s\boldsymbol{p}_i = \begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \boldsymbol{C}[\boldsymbol{R} \quad \boldsymbol{t}]_{4\times4}\boldsymbol{P} = [...]_{3\times3} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \tag{22}$$

where $\boldsymbol{p}_i$ is the projective image point, s an arbitrary scale factor, and $\boldsymbol{C}$ and $\boldsymbol{R} \ \ \boldsymbol{t}$ are the camera intrinsic and extrinsic parameters.

From the above equation, we can see that there is a fixed mapping between ground plane points and image points. This mapping is called a collineation or planar homography of points. A ground plane point is related to the camera image by a collineation $\boldsymbol{H_c}$:

$$s\boldsymbol{p}_i = \boldsymbol{H_c}.\widetilde{\boldsymbol{P}}, \tag{23}$$

where $\widetilde{\boldsymbol{P}} = (X, Y, 1)^\top$ and

$$\boldsymbol{H_c} = \boldsymbol{C}[\boldsymbol{r}_1 \quad \boldsymbol{r}_2 \quad \boldsymbol{t}], \tag{24}$$

with $\boldsymbol{r_i}$ denoting the $i\mathrm{th}$ column of the rotation matrix $\boldsymbol{R}$.

For a stereo system, we can express the collineation between ground plane points and the left and right cameras

$$s\boldsymbol{p}_{li} = \boldsymbol{H_l}.\widetilde{\boldsymbol{P}} \ \ and \ \ s\boldsymbol{p}_{ri} = \boldsymbol{H_r}.\widetilde{\boldsymbol{P}}, \tag{25}$$

where $\boldsymbol{p}_{li}$ and $\boldsymbol{p}_{ri}$ are the left and right projective image points.

We can consider a direct mapping $\boldsymbol{H}$ of ground plane points between the stereo pair. $\boldsymbol{H}$ can be obtained by calibration using known ground plane points [41], or using (24) and known camera intrinsic and extrinsic parameters $\boldsymbol{C}$ and $\boldsymbol{R}$ $\boldsymbol{t}$. For the direct mapping $\boldsymbol{H}$ of right image points to the left image, we have

$$s\boldsymbol{p}_{li} = \boldsymbol{H}.\boldsymbol{p}_{ri} = \boldsymbol{H_l}.\boldsymbol{H_r}^{-1}.\boldsymbol{p}_{ri}. \tag{26}$$

To obtain $\boldsymbol{H}$, we must first compute $\boldsymbol{H_l}$ and $\boldsymbol{H_r}$. From (24) and using ${}^{\mathcal{C}}\boldsymbol{T}_{\mathcal{N}}$ obtained from the inertial data and ${}^{\mathcal{L}}\boldsymbol{T}_{\mathcal{C}}$ obtained form the geometric setup, $\boldsymbol{H_l}$ is given by

$$\begin{aligned}\boldsymbol{H_l} &= \boldsymbol{C_L}[\boldsymbol{r}_1 \quad \boldsymbol{r}_2 \quad \boldsymbol{t}]_L \\ &= \boldsymbol{C_L}\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\end{bmatrix}.{}^{\mathcal{L}}\boldsymbol{T}_{\mathcal{C}}.{}^{\mathcal{C}}\boldsymbol{T}_{\mathcal{N}}.\begin{bmatrix}1&0&0\\0&1&0\\0&0&0\\0&0&1\end{bmatrix},\end{aligned} \tag{27}$$

and proceeding analogously for the right camera, we obtain $\boldsymbol{H_r}$.

From (26), we have that the collineation between left and right images of ground plane points, for a system with symmetric verge angle θ and baseline b, is given by

$$\boldsymbol{H} = \boldsymbol{H_l}.\boldsymbol{H_r}^{-1} = \begin{bmatrix} -\frac{2n_x b\cos\theta\sin\theta+n_y b+2d-4d\cos^2\theta}{-n_y b+2d} & \frac{2bn_z\cos\theta}{-n_y b+2d} & f\frac{-2\cos\theta(2d\sin\theta+bn_x\cos\theta)}{-n_y b+2d} \\ 0 & 1 & 0 \\ 2\frac{2d\sin\theta\cos\theta-n_x b+n_x b\cos^2\theta}{f(-n_y b+2d)} & \frac{2bn_z\sin\theta}{f(-n_y b+2d)} & -\frac{2n_x b\cos\theta\sin\theta+n_y b-4d\cos^2\theta+2d}{-n_y b+2d} \end{bmatrix}, \tag{28}$$

where f is the camera focal distance, $(n_x, n_y, n_z)^\top$ the vertical reference provided by the inertial sensors given in the camera system frame of reference $\{\mathcal{C}\}$ (with origin at the middle of the baseline), and d the system height to the ground plane. This equation will be fundamental for the world feature detection methods described in following sections.

This collineation can also be computed for other planes. Consider a mobile robot going up a slope. In this case, the leveled ground plane is no longer relevant, but we can consider the local planar patch with normal $\boldsymbol{n}_s \neq \boldsymbol{n}$ given by the robots steady state tilt and proceed as before.

6 Case Studies Using the Vertical Reference

6.1 Focal Distance Calibration

Camera calibration using vanishing points has been widely explored, [31], [42], [43], [44], [45] among others. The novelty in our work is using just one vanishing point and using the inertial sensors to extract camera pose information. Calibration based on vanishing points is limited since a compromise has to be reached on the quality of each point, but since we require just one vanishing point, the best one can be chosen.

Vanishing point $\boldsymbol{p}_v = (u, v)^\top$, obtained from a set of parallel lines belonging to some leveled plane, and $\hat{\boldsymbol{n}} = (n_x, n_y, n_z)^\top$ taken from (9), are conjugate to each other since they correspond to 3D orthogonal sets of parallel lines. From (16), the focal distance f can be estimated as

$$f = -\frac{n_x u + n_y v}{n_z}. \tag{29}$$

With a suitable calibration target scene where ground plane parallel lines can be easily found, the focal distance can be estimated using (29). The image center is assumed to be fixed and known, and (u, v) are given in image centered coordinates. If no prior calibration is done to determine the image center, nonimaging techniques, such as numerical center of image or sensor coordinates, are used. The implications of this assumption depend on the camera quality and variable parameters [46].

The orthogonality of two leveled plane sets of parallel lines, when using two vanishing points, is replaced here by the orthogonality between vertical lines, with vanishing point $(n_x, n_y, n_z)^\top$, and a set of leveled parallel lines, with vanishing point $(u, v, f)^\top$. This implies that the alignment between the IMU and the camera has to be known from construction or calibration.

The effect of errors in the vertical reference on the estimated focal distance can be seen by studying the Jacobian matrix

$$\begin{aligned}\boldsymbol{J} &= \begin{bmatrix}\frac{\partial f}{\partial n_x} & \frac{\partial f}{\partial n_y} & \frac{\partial f}{\partial n_z} & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v}\end{bmatrix} \\ &= \begin{bmatrix}-\frac{u}{n_z} & -\frac{v}{n_z} & \frac{n_x u+n_y v}{n_z^2} & -\frac{n_x}{n_z} & -\frac{n_y}{n_z}\end{bmatrix}.\end{aligned} \tag{30}$$

Considering a good pose tilting down about $45\,deg$ with a not too distant vanishing point, having $\hat{\boldsymbol{n}} = (0, -0.70, 0.71)^\top$ and $(u, v)^\top = (100, 1000)^\top$, a $1\,deg$ error in the vertical reference would perturb the estimated focal distance f by $\pm\Delta f$ with

$$\Delta f = \sin(1^\circ)\frac{n_x u + n_y v}{n_z^2} \approx 24.2. \tag{31}$$

This means a 2.5 percent error in the estimated value of $f \approx 986$. This error will degrade when more ill conditioned poses are used, and the solution degenerates when observability of a ground plane vanishing point is missing or the camera is perfectly horizontal, with the horizon line through the image center.

6.2 Stereo Correspondence of Ground Plane Points and 3D Position

Since we know the collineation of the ground plane image points from (28), image points can be tested across the stereo pair, identifying the ground plane points, and determining their 3D position. An algorithm for the 3D reconstruction of image detected features can be formulated. For each detected point in the right image $\boldsymbol{p}_{r_i}$, map it to the other image using the known collineation. The correspondent point in the left image is found by parsing all the left image detected points of interest $\boldsymbol{p}_{l_j}$ and testing an allowed neighbourhood window for a match, i.e., find j such that

$$\boldsymbol{p}_{l_j} = \boldsymbol{H}.\boldsymbol{p}_{r_i} \pm \boldsymbol{\delta}. \qquad (32)$$

If there is a match, the point belongs to the ground plane. If there is no match, the point must be something other than the floor, possibly an obstacle. If the detected interest points are very dense, false positives will occur since it will be easy to have some other point in the same neighborhood. To overcome this, 2D correlation is performed over a small region around both image points.

From (25), the 3D position ${}^{\mathcal{N}}\boldsymbol{P} = (X, Y, 0, 1)^\top$ of this ground plane point is given by

$${}^{\mathcal{N}}\widetilde{\boldsymbol{P}} = \boldsymbol{H_r}^{-1}\boldsymbol{p}_{r_i}, \qquad (33)$$

where ${}^{\mathcal{N}}\widetilde{\boldsymbol{P}} = (X, Y, 1)^\top$.

Errors in the estimated vertical reference will increase the uncertainty, but since the method relies on a neighborhood test, it maintains robustness in detecting points up to the tolerance of the used search window size and then breaks down. The 3D mapping error, however, will degrade with increasing error in the vertical reference. A statistical map of the detected features has to be built to deal with the uncertainty.

6.3 Image Line Segmentation

Knowing the vertical, the vanishing point of all image lines that correspond to world vertical features is known. This vanishing point is at infinity when there is no tilt, and the vertical lines are all parallel in the image. For small tilt values, the vertical lines can be taken as parallel, speeding up the detection process. Based on this assumption, the vertical line segments found in the image will be parallel to the local image vertical $\hat{\boldsymbol{n}}_i$, the normalized image projection of the vertical $\hat{\boldsymbol{n}}$. The image vertical reference corresponds to the unit sphere projection of the vanishing point of all 3D vertical lines in the image plane.

In order to detect vertical lines, we extracted the edges in the image using a modified Sobel filter [47]. By choosing an appropriate threshold for the gradient magnitude, the potential edge lines can be identified. The square of the gradient was used in our application to allow faster integer computation.

To only obtain the vertical edges we compare the pixel gradient with the vertical. The dot product of the gradient with the vertical should be null, so by setting a tolerance threshold value, the detected edge points can be taken as vertical or not.

$$\boldsymbol{\mathcal{D}}.\hat{\boldsymbol{n}}_i < tolerance. \qquad (34)$$

But, this can lead to erroneous results since the pixel gradient provides a very local information and is affected by the pixel quantization, therefore, a large tolerance is used. In order to extract the vertical lines in the image, all edge points that satisfied (34) were mapped to a rectified image table, using (35), so that continuity could be tested along the vertical edge direction. So, each edge point $\boldsymbol{p}_j = (u, v)$ contributed to the table at position

$$vert\_points\ (x, y) = \left(\boldsymbol{p}_j.\hat{\boldsymbol{h}}_i, \boldsymbol{p}_j.\hat{\boldsymbol{n}}_i\right), \qquad (35)$$

where $\hat{\boldsymbol{h}}_i$ is the horizontal unit vector, perpendicular to $\hat{\boldsymbol{n}}_i$ in the image plane, i.e.,

$$\hat{\boldsymbol{n}}_i.\hat{\boldsymbol{h}}_i = 0. \qquad (36)$$

The minimum line length and allowable gaps is set and each column of the table parsed. The end result is a set of image lines, given by their end-points in the original image, that correspond to 3D vertical features, except for degenerate cases for which multiple views are required.

For large tilt values, the vertical lines can not be taken as parallel, and must be tested to comply with vanishing point $\hat{\boldsymbol{n}}$. If $\boldsymbol{m}$ is the unit vector normal to the line projection plane, the 3D line can only be vertical if

$$\hat{\boldsymbol{n}}.\boldsymbol{m} = 0, \qquad (37)$$

but, again, with a single view, a false vertical might be detected in degenerate cases.

6.4 Stereo Correspondence of Vertical Lines and 3D Position

In the previous section, a method was presented for vertical image line detection. But, in order to have world feature detection, the image segmentation of vertical lines has to be matched across the stereo pair, and the 3D position of the feature determined.

Making the assumption that the relevant vertical features start from the ground plane, and since we know the collineation of the ground plane image points from (28), a common unique point is identified. The lower point or *foot* of each vertical feature in one image should map to the corresponding *foot* in the other image.

Proceeding as before, the *feet* of the vertical line features can be tested across the stereo pair using the known collination. If there is a match, the point belongs to the ground plane and must be the *foot* of a true 3D vertical world feature. The 3D position of the *foot* of this vertical element is given by (33).

With the system mounted on some mobile robot, the vertical features can be charted on a world map, constructed as the robot moves in its environment. This map is constructed in the robot's navigation frame of reference $\{\mathcal{N}\}$ as described in Section 4.7.

Besides the error in detected points and their 3D mapping previously mentioned, the vertical edge detection also used the vertical reference, but only as a rough estimate. Under the assumption that near vertical features are rare, this does not present a problem.

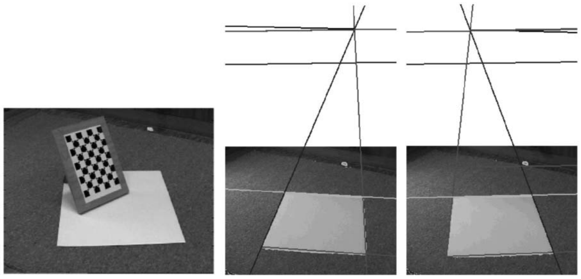


Fig. 8. One of the 20 images used in the calibration and estimation of f at two target positions with a near vanishing point, showing horizon line with initial guess value of f (lower) and correct horizon line given by inertial vertical reference (top).

7 Results

Fig. 7 shows an inertial system prototype built at our lab [25] that was coupled to a camera stereo rig to carry out the tests.

7.1 Focal Distance Calibration

To test the estimation of f using one vanishing point and the vertical reference, the camera calibration toolbox provided by Intel Open Source Computer Vision Library [48] was used to provide a standard camera calibration method. The calibration used images of a chessboard target in several positions and recovers the camera's intrinsic parameters as well as the target positions relative to the camera. The calibration algorithm is based on Zhang's work in estimation of planar homographies for camera calibration [49], but the closed-form estimation of the internal parameters from the homographies is slightly different, since the orthogonality of vanishing points is explicitly used and the distortion coefficients are not estimated at the initialization phase.

The calibration was performed with 20 images of a chessboard target in several positions, as seen in Fig. 8. Without changing the camera, the chessboard target was removed and the calibration was performed using just one vanishing point and the inertial vertical reference. Two target positions with a near vanishing point were used, as seen in Fig. 8, and 100 samples taken at each position. From Fig. 9 and Table 2, we can see that the proposed method provides a good estimate of f, within the uncertainty of the standard method used.

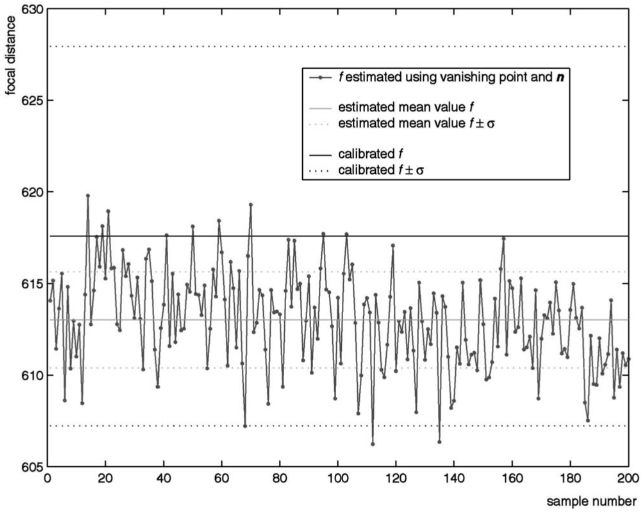


Fig. 9. Estimation of f with just one vanishing point and $\hat{n}$, compared with camera calibration results. Two target positions with a near vanishing point, 100 samples taken at each position.

TABLE 2
Estimation of f

	mean	σ
20 images of chessboard target	617.57	10.36
$\hat{n}$ & vanishing point	613.02	2.62

The main sources of error are the vanishing point instability, evidenced by the stepwise results obtained in other tests [29], and the noise in the vertical reference provided by the low cost accelerometers. The results show that the proposed method is feasible. Due to its simplicity, it can be performed on-the-fly by a mobile robot in a man made environment, where ground plane parallel lines can be easily detected. It can also aid 3D modeling and reconstruction by providing extra information about focal distance when digitally acquiring an image, as in [21].

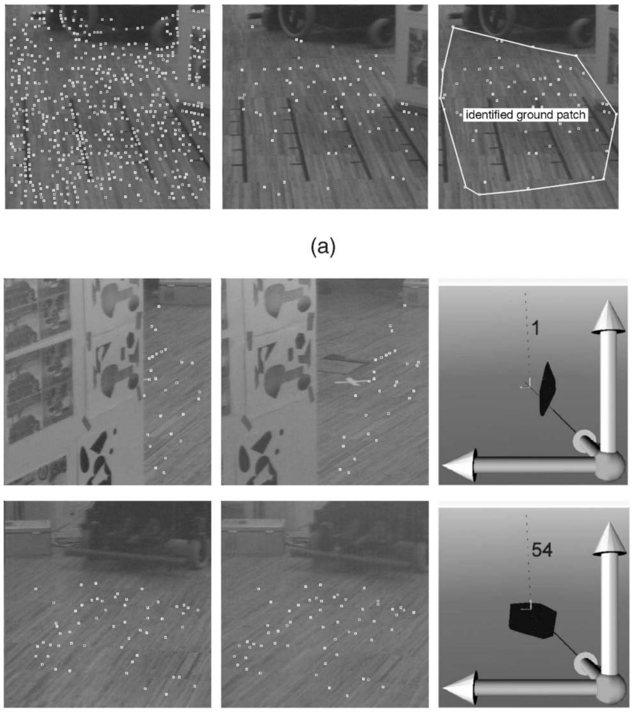
7.2 Ground Plane Segmentation

The ground plane segmentation algorithm was implemented with a previous version of our system mounted on a mobile robot. The points shown in Fig. 10 were obtained using SUSAN [50] corner detector. The points of interest in the right image were then parsed as described in the previous section. Grahams Algorithm [51] was used for computation of the convex polygon involving the set of points. Fig. 10 also shows some frames from a ground plane detection sequence obtained with the system on a mobile robot and corresponding VRML view of the ground patch.

For visualization of the detected ground points, a VRML world was generated [52]. The identified ground plane patch was mapped onto the 3D scene, as seen in Fig. 10. The complete sequence was processed, generating polygons corresponding to the identified ground plane patch for each frame. To update the VRML world on-the-fly, only the ground patch vertex points need to be sent, so that the polygon can be rendered. When bandwidth is not a problem, the segmented image patch can also be sent and placed onto the polygon. VRML opens many other possibilities such as teleoperation or path-planning environments.

Adjusting a convex polygon to the set of points can lead to erroneous ground patch segmentation. Some changes have to be made to the algorithm and special cases taken into account, such as having multiple isolated polygons or allowing for nonconvexity when points are too far apart and an obstacle could be in the way.

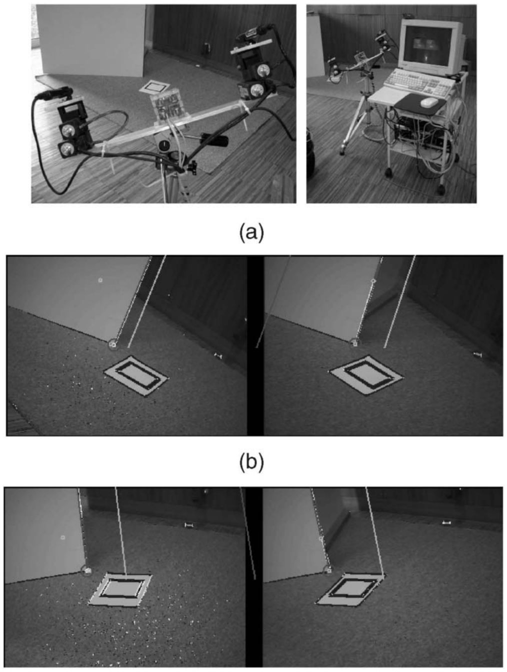
The results show that the method works, but is very dependant on texture so that feature points can be detected. There are many initial feature points, but only a few are correctly detected as ground plane points, with many false negatives. If, instead of detecting the ground plane, an



(a)

(b)

Fig. 10. (a) Right image with a set of initial points, detected ground plane points, and identified ground patch. (b) Stereo frames from ground plane detection sequence with the VRML view of the ground patch on the right side.



(a)

(b)

(c)

Fig. 11. (a) The experimental setup. (b) and (c) Vertical world feature detection. The bigger circles indicate the *foot* of a detected vertical world feature, the smaller circles the points tested, i.e., the lower end of image vertical lines.

obstacle detection was being done, these unmatched points could be perceived as obstacles. This can be avoided by making assumptions on the minimum size of obstacles and detected point density. This method enables fast processing of images and feature matching across the stereo pair, since the ground plane restriction is used to limit the search space.

7.3 Detecting Vertical Lines and 3D Position

We implemented the vertical world feature detector with our system, working real-time at five frames per second. An initial setup had to be done to properly align the cameras and verge them with a known angle, using the pan and tilt units.

Fig. 11 shows a set of results. The system was initially fixated on a ground plane point, using the pan and tilt units to verge with a known angle, so that system height could be determined. Keeping a constant height, the system was tilted sideways, and the vertical feature was correctly detected in all frames. Further tests showed that method performs well in man made environments where vertical features are abundant, but required some parameter adjustment to have good results with different types of scenes.

Using (33) and (18), the vertical features are then charted on a world map. Fig. 12 shows the output of the vertical world feature detector that includes a map with detected features. The system was placed on a mobile robot and placed at the entry hall of our lab. The maps shows the furniture correctly mapped. The raw data shows a spread along the line of sight of the system, as expected from the geometric setup and image noise. Proper time filtering and outlier removal has to be performed to have a consistent map. The map has to be updated as the robot moves in its environment, but this was not implemented in this work. Part of this work was presented at [28] and ongoing work at our lab is being done in the mapping [53].

8 Discussion and Conclusions

This article sets a framework for inertial and visual sensor cooperation and presented some results of using gravity as a vertical reference. From the studies on human vision, it is clear that inertial cues play an important role and that the notion of vertical is important at the first stages of image processing. Further studies in the field, as well as bioinspired robotic applications, will enable a better understanding of the underlying principles, with possible application for bioimplants of artificial vision and vestibular systems in patients.

The unit sphere projection model used provides an intuitive representation of projective geometry, onto which inertial cues are easily integrated. Exploring the orthogonality between the vertical reference and vanishing points of horizontal lines, camera focal distance was estimated using only one vanishing point. This allows the best vanishing point to be chosen, and is less imposing on the availability of scene vanishing points. An integrated accelerometer and imaging sensor could use this method to estimate focal distance, relying on the automatic detection of one vanishing point of a set of horizontal lines, with high probability of

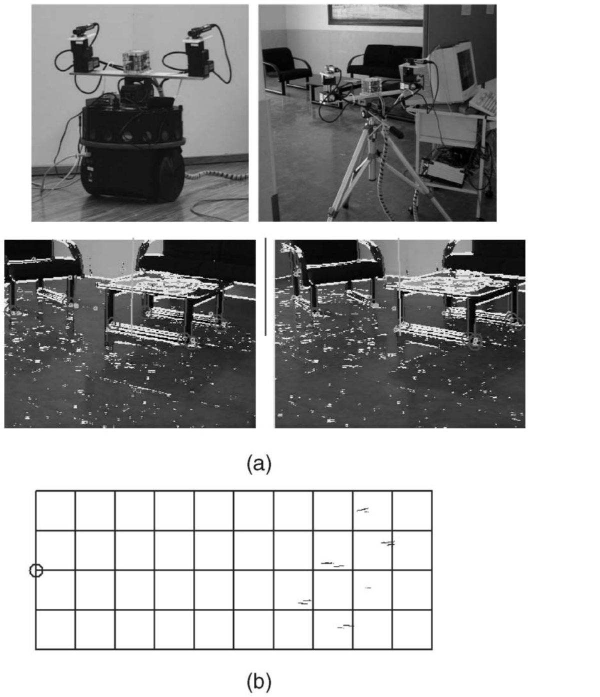


Fig. 12. (a) The experimental setup with the system placed on a mobile robot and placed at the entry hall of our lab. (b) Vertical world feature detection. The circle in the map represents the robot and the points the detected vertical world features.

occurring at specific camera poses in man-made structured environments. When applied to mobile robots, the vanishing point can also provide an external bearing for the navigation frame of reference. Calibration methods using specific calibration targets and multiple images can provide more precise focal distance estimates. The main sources of error in this method are the uncertainty in the vanishing point estimation, the assumed alignment of the inertial sensors, and the accelerometer noise.

In a stereo rig with known geometry, the vertical reference was used to compute the collineation of level plane points, enabling their detection and 3D mapping. This was used to segment and reconstruct vertical features and leveled planar patches. These 3D world features are useful to improve mobile robot autonomy and navigation. The method is fast and adaptable, unlike a fixed calibrated collineation estimated from a set of known points. The main sources of error in this method are the assumed known geometry and the noise in the vertical reference. When used on a mobile vehicle the error increased along the direction of motion [53], but still provided a useful map of vertical features for robot navigation, where the uncertainty can be modeled.

Another approach we followed, not covered in this paper, was to use standard vision techniques to compute depth maps and then rotate and align them using the inertial reference [54], [55]. The advantage of reducing the search space explored above is lost, but current technology provides real-time depth maps with reasonable quality, and the inertial data fusion is still very useful at a later step to align and register the obtained maps.

References

[1] H. Carpenter, *Movements of the Eyes.* second ed. London Pion Limited, 1988.

[2] A. Berthoz, *The Brain's Sense of Movement.* Harvard Univ. Press, 2000.

[3] K.K. Gillingham and F.H. Previc, *Spatial Orientation in Flight.* second ed. chapter 11, Williams and Wilkins, 1996.

[4] R.O. Eason and R.C. Gonzalez, "Least-Squares Fusion of Multisensory Data," *Data Fusion in Robotics and Machine Intelligence,* M.A. Abidi and R.C. Gonzalez, eds., chapter 9, Academic Press, 1992.

[5] R.P.G. Collinson, *Introduction to Avionics.* Chapman & Hall, 1996.

[6] G.R. Pitman, *Inertial Guidance.* John Wiley & Sons, 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," *Proc. Position Location and Navigation Symposium,* Apr. 1998.

[8] A.C. Luther, *Video Camera Technology.* Artech House Publishers, 1998.

[9] T. Viéville and O.D. Faugeras, "Computation of Inertial Information on a Robot," *Proc. Fifth Int'l Symp. Robotics Research,* pp. 57-65, H. Miura and S. Arimoto, eds., MIT Press, 1989.

[10] T. Viéville and O.D. Faugeras, "Cooperation of the Inertial and Visual Systems," *Traditional and NonTraditional Robotic Sensors,* pp. 339-350. T.C. Henderson, ed., Springer Verlag, 1990.

[11] 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," *Intelligent Robots and Systems,* M. Kikode, T. Sato, and K. Tatsuno, eds., 1993.

[12] T. Viéville, E. Clergue, and P.E.D. Facao, "Computation of Ego-Motion and Structure from Visual and Inertial Sensor Using the Vertical Cue," *Proc. Int'l Conf. Computer Vision,* pp. 591-598, 1993.

[13] T. Viéville, *A Few Steps Towards 3D Active Vision.* Springer-Verlag, 1997.

[14] B. Bhanu, B. Roberts, and J. Ming, "Inertial Navigation Sensor Integrated Motion Analysis for Obstacle Detection," *Proc. IEEE Int'l Conf. Robotics and Automation,* pp. 954-959, 1990.

[15] F. Panerai and G. Sandini, "Visual and Inertial Integration for Gaze Stabilization," *Proc. Int'l Symp. Intelligent Robotic Systems,* 1997.

[16] F. Panerai and G. Sandini, "Oculo-Motor Stabilization Reflexes: Integration of Inertial and Visual Information," *Neural Networks,* vol. 11, nos. 7-8, pp. 1191-1204, 1998.

[17] F. Panerai, G. Metta, and G. Sandini, "Visuo-Inertial Stabilization in Space-Variant Binocular Systems," *Robotics and Autonomous Systems,* vol. 30, nos. 1-2, pp. 195-214, 2000.

[18] 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," *Proc. Seventh Int'l Conf. Computer Vision,* pp. 411-417, Sept. 1999.

[19] T. Mukai and N. Ohnishi, "Object Shape and Camera Motion Recovery Using Sensor Fusion of a Video Camera and a Gyro Sensor," *Information Fusion,* vol. 1, no. 1, pp. 45-53, 2000.

[20] R. Kurazume and S. Hirose, "Development of Image Stabilization System for Remote Operation of Walking Robots," *Proc. IEEE Int'l Conf. Robotics and Automation,* pp. 1856-1860, Apr. 2000.

[21] S.R. Coorg, "Pose Imagery and Automated Three-Dimensional Modeling of Urban Environments," PhD thesis, Massachusetts Inst. of Technology, Sept. 1998.

[22] S. You, U. Neumann, and R. Azuma, "Hybrid Inertial and Vision Tracking for Augmented Reality Registration," *Proc. IEEE Virtual Realiy Conf.,* pp. 260-267, Mar. 1999.

[23] W.A. Hoff, K. Nguyen, and T. Lyon, "Computer Vision-Based Registration Techniques for Augmented Reality," *Proc. Conf. Intelligent Robots and Computer Vision,* pp. 538-548, Nov. 1996.

[24] E.D. Dickmanns, "Vehicles Capable of Dynamic Vision: A New Breed of Technical Beings?" *Artificial Intelligence,* vol. 103, pp. 49-76, 1998.

[25] J. Lobo and J. Dias, "Integration of Inertial Information with Vision towards Robot Autonomy," *Proc. IEEE Int'l Symp. Industrial Electronics,* pp. 825-830, July 1997.

[26] J. Lobo, L. Marques, J. Dias, U. Dias, and A.T. de Almeida, *Sensors for Mobile Robot Navigation,* pp. 50-81. Springer-Verlag, 1998.

[27] J. Lobo and J. Dias, "Ground Plane Detection Using Visual and Inertial Data Fusion," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems,* pp. 912-917, Oct. 1998.

[28] J. Lobo, C. Queiroz, and J. Dias, "Vertical World Feature Detection and Mapping Using Stereo Vision and Accelerometers," *Proc. Ninth Int'l Symp. Intelligent Robotic Systems,* pp. 229-238, July 2001.

[29] J. Lobo and J. Dias, "Fusing of Image and Inertial Sensing for Camera Calibration," *Proc. Int'l Conf. Multisensor Fusion and Integration for Intelligent Systems,* pp. 103-108, Aug. 2001.

[30] J. Lobo, "Inertial Sensor Data Integration in Computer Vision Systems," MS thesis, Univ. of Coimbra, Apr. 2002.

[31] K. Kanatani, *Geometric Computation for Machine Vision.* Oxford Univ. Press, 1993.

[32] M.D. Shuster, "The Kinematic Equation for the Rotation Vector," *IEEE Trans. Aerospace and Electronic Systems,* vol. 29, no. 1, pp. 263-267, Jan. 1993.

[33] P.G. Savage, *Strapdown System Algorithms.* chapter 3, pp. 1-30, AGARD: Advisory Group for Aerospace Research and Development, 1984.

[34] M.J. Caruso, T. Bratland, C.H. Smith, and R. Schneider, "A New Perspective on Magnetic Field Sensing," technical report, Honeywell, Inc., 1998.

[35] Summit Instruments, http://www.summitinstruments.com, 2003.

[36] Analog Devices, Mems Integrated Micro-Electromechanical Systems, Analog Devices, iMEMS @ http://www.analog.com/, 2003.

[37] B.K.P Horn, "Closed-Form Solution of Absolute Orientation Using Unit Quaternions," *J. Optical Soc. Am.,* vol. 4, no. 4, pp. 629-462, Apr. 1987.

[38] J. Alves, J. Lobo, and J. Dias, "Camera-Inertial Sensor Modelling and Alignment for Visual Navigation," *Proc. 11th Int'l Conf. Advanced Robotics,* pp. 1693-1698, July 2003.

[39] J. Dias, C. Paredes, I. Fonseca, and A.T. de Almeida, "Simulating Pursuit with Machines," *Proc. IEEE Conf. Robotics and Automation,* pp. 472-477, 1995.

[40] J. Dias, C. Paredes, I. Fonseca, H. Araujo, J. Baptista, and A.T. de Almeida, "Simulating Pursuit with Machine Experiments with Robots and Artificial Vision," *IEEE Trans. Robotics and Automation,* vol. 3, no. 1, pp. 1-18, Feb. 1998.

[41] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge Univ. Press, 2000.

[42] L.-L. Wang and W.-H. Tsai, "Camera Calibration by Vanishing Lines for 3-D Computer Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 4, pp. 370-376, Apr. 1991.

[43] B. Caprile and V. Torre, "Using Vanishing Points for Camera Calibration," *Proc. Int'l J. Computer Vision,* vol. 4, no. 2, pp. 127-140, 1990.

[44] B. Brillault and B. O'Mahony, "New Method for Vanishing Point Detection," *Computer Vision, Graphics and Image Processing: Image Understanding,* vol. 54, no. 2, pp. 289-300, 1991.

[45] M. Li, "Camera Calibration of the Kth Head-Eye System," *Proc. European Conf. Computer Vision,* pp. 543-554, 1994.

[46] R.G. Willson and S.A. Shafer, "What is the Center of the Image," *J. Optical Soc. Am. A,* vol. 11, no. 11, pp. 2946-2955, 1994.

[47] B. Jahne, *Digital Image Processing.* Springer-Verlag, 1997.

[48] Intel, Intel Open Source Computer Vision Library, http://www.intel.com/research/mrl/research/opencv/, 2003.

[49] Z. Zhang, "Flexible Camera Calibration By Viewing a Plane From Unknown Orientations," *Proc. Seventh Int'l Conf. Computer Vision,* pp. 666-673, Sept. 1999.

[50] S.M. Smith and J.M. Brady, "SUSAN—A New Approach to Low Level Image Processing," *Int'l J. Computer Vision,* vol. 23, no. 1, pp. 45-78, May 1997.

[51] J. O'Rourke, *Computational Geometry in C.* Cambridge Univ. Press, 1993.

[52] A.L. Ames, D.R. Nadeau, and J.L. Moreland, *VRML 2.0 Sourcebook.* second ed. John Wiley and Sons, 1997.

[53] J. Lobo, C. Queiroz, and J. Dias, "World Feature Detection and Mapping Using Stereovision and Inertial Sensors," *Robotics and Autonomous Systems,* vol. 44, no. 1, pp. 69-81, July 2003.

[54] J. Lobo, L. Almeida, and J. Dias, "Segmentation of Dense Depth Maps Using Inertial Data: A Real-Time Implementation," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems,* pp. 92-97, Oct. 2002.

[55] J. Lobo and J. Dias, "Inertial Sensed Ego-Motion for 3D Vision," *Proc. InerVis Workshop, 11th Int'l Conf. Advanced Robotics,* pp. 1907-1914, July 2003.

Jorge Lobo completed his five year course in electrical engineering at Coimbra University in 1995. He was a junior teacher in the Computer Science Department of the Coimbra Polytechnic School, and later joined the Electrical and Computer Engineering Department of the Faculty of Science and Technology at the University of Coimbra, where he currently works as a teaching assistant. In April 2002, he received the MSc degree with the thesis "Inertial Sensor Data Integration in Computer Vision Systems." His current research is carried out at the Institute of Systems and Robotics, University of Coimbra, and is aimed at the fusion of inertial information with vision systems in mobile robots.

Jorge Dias received the electrical engineer degree (specialization on computers) from the Faculty of Science and Technology from the University of Coimbra in July 1984. He received the PhD degree in electrical engineering from the University of Coimbra with specialization in control and instrumentation, in November 1994. His main research area is computer vision and robotics, with activities and contributions on the field since 1984. He has been exploring different topics on computer vision and mobile robotics toward the improvement of autonomous robotic systems. He was the main researcher for projects financed by the Portuguese Institution JNICT—Junta Nacional de Investigao Cientfica, FCT—Portuguese Foundation for Science and Technology. He worked as an investigator in projects financed by the European Community and NATO—Science For Stability Program.