An Introduction to Inertial and Visual Sensing

Peter Corke, Jorge Lobo and Jorge Dias

The International Journal of Robotics Research 2007 26: 519
DOI: 10.1177/0278364907079279

Peter Corke
CSIRO ICT Centre
Brisbane, Australia
peter.corke@csiro.au

Jorge Lobo*
Jorge Dias
Institute of Systems and Robotics
University of Coimbra, Portugal
{jlobo,jorge}@isr.uc.pt

An Introduction to Inertial and Visual Sensing

Abstract

In this paper we present a tutorial introduction to two important senses for biological and robotic systems — inertial and visual perception. We discuss the fundamentals of these two sensing modalities from a biological and an engineering perspective. Digital camera chips and micro-machined accelerometers and gyroscopes are now commodities, and when combined with today's available computing can provide robust estimates of self-motion as well 3D scene structure, without external infrastructure. We discuss the complementarity of these sensors, describe some fundamental approaches to fusing their outputs and survey the field.

KEY WORDS—vision, inertial sensing, sensor fusion

1. Introduction

All animals make use of multiple sensory modalities. As children we learn about the five senses: vision, smell, hearing, touch and taste but in fact we have many more, including joint position, muscle exertion, balance and motion. Some animals (Hughes 1999) have developed specialized sensors for acoustic ranging (echo-location in bats), magnetic dead-reckoning (navigation using magnetic fields in some birds) and detection of prey by incredibly sensitive detection of electric fields (some sharks and eels). We also combine some of our sensing modalities; balance and motion from the inner ear, joint position and vision into a virtual sense of movement which is called *kinesthesia*.

For mobile robotics, ground, air and underwater, a sense of position (localization) and motion are critically important. The senses and fusion techniques evolved by animals may help us to achieve a level of robotic competency that matches or exceeds that of animals. In robotics we have available to us sensors that have no biological analog, for example GPS (Global Positioning System), radar and LIDAR (LIght Detection And Ranging).

In this paper we will discuss kinesthesia for robotics – how to integrate information from vision and inertial sensors to provide a robust and non-ambiguous representation of robotic motion. We will cover the fundamentals of these two sensing modalities from the perspectives of physical principles and the engineering and biological implementations. We will show that these sensors have useful complementarities, each able to cover the limitations and deficiencies of the other. From an engineering perspective this is extremely useful, and that nature has found it useful to evolve such a complementary sensing system is interesting and compelling.

A useful way to consider sensors is in terms of the spatial derivative that they sense. GPS and vision are both able to sense actual position with order 0, while odometry and gyroscopes sense order 1 (translational and rotational velocity) and accelerometers sense order 2 (translational acceleration). Higher order derivatives have the advantage of rapidly sensing the onset of motion but their integration over time can lead to unbounded errors if offsets and scale errors are present. However while GPS seems ideal and is a very common and low-cost sensor it has many limitations. Standard GPS has a substantial error (of order 10 m) when used without differential or RTK (Real Time Kinematic) correction, and requires line of sight to the satellite constellation which rules out operation underwater, underground, in many urban environments and even beneath dense tree cover.

The International Journal of Robotics Research
Vol. 26, No. 6, June 2007, pp. 519–535
DOI: 10.1177/0278364907079279

Figures 2, 6, 7 appear in color online: http://ijr.sagepub.com

* Corresponding author

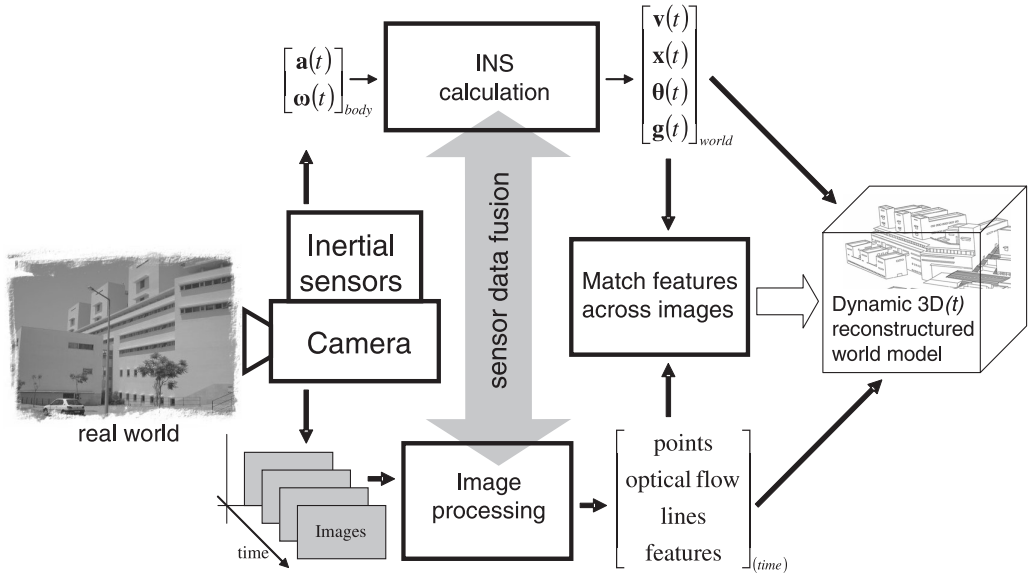

Fig. 1. Combining inertial and vision sensing.

A single vision sensor can measure relative position with derivative order of 0 but senses only a 2D projection of the 3D world – direct depth information is lost but can be inferred, for example using stereo vision or integration of data from multiple viewpoints. Optical flow can be computed numerically from an image sequence to provide derivative order 1 information, object velocity scaled by distance which may not be known. Inertial sensors such as gyroscopes and accelerometers measure derivative order 1 (angular velocity) and order 2 (translational acceleration) respectively. If information from vision and inertial sensors can be combined, spanning derivative orders from 0 to 2, the result would be a very useful robot sensor, particularly since it would require no external infrastructure.

Figure 1 shows a framework for combination of inertial and vision sensors. The 3D world is observed by the visual sensor while its pose and motion parameters are estimated by the inertial sensors. These motion parameters can also be inferred from the image flow and known scene features. 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. Pure structure from motion (Hartley and Zisserman 2004) and bundle-adjustment (Triggs et al. 2000) methods can achieve similar results but at much higher computational cost and with lower robustness.

Today both types of sensor are low-cost, high-performance and becoming commodities. Color CMOS cameras are mass produced for mobile phones and some of these have zoom and focus control. Micro-machined (MEMS) accelerometers have long been used for automotive air-bag triggers and are now used in laptops to detect free-fall, and in digital cameras to sense camera orientation.

The remainder of this tutorial is organized as follows. Section 2 describes the fundamentals of inertial sensing and section 3 covers visual sensing. Section 4 describes the principles behind fusion of these two senses and applications.

2. Inertial Sensing

Gyroscopes and accelerometers are known as inertial sensors since they exploit the property of inertia, i.e. resistance to a change in momentum, to sense angular motion in the case of the gyro, and changes in linear motion in the case of the accelerometer. Inclinometers are also inertial sensors and measure the orientation of the acceleration vector due to gravity. Inertial sensors are not dependent on any external references or infrastructure, apart from the ubiquitous gravity field.

An inertial measurement unit (IMU) typically comprises three orthogonal accelerometers to measure the acceleration of the body, and also include three orthogonal gyroscopes to measure the rate of change of the body's orientation. Linear velocity and position, and angular position are obtained by integration. This is the principle behind inertial navigation systems (INS) which are used in aerospace and naval applications (Lawrence 1998). Over the last 15 years the developments in electronic and silicon micromachining, pushed by the needs of the automotive and consumer industry, have brought about low-cost batch fabricated, silicon sensors (Yazdi et al. 1998), which in turn is leading to new applications.

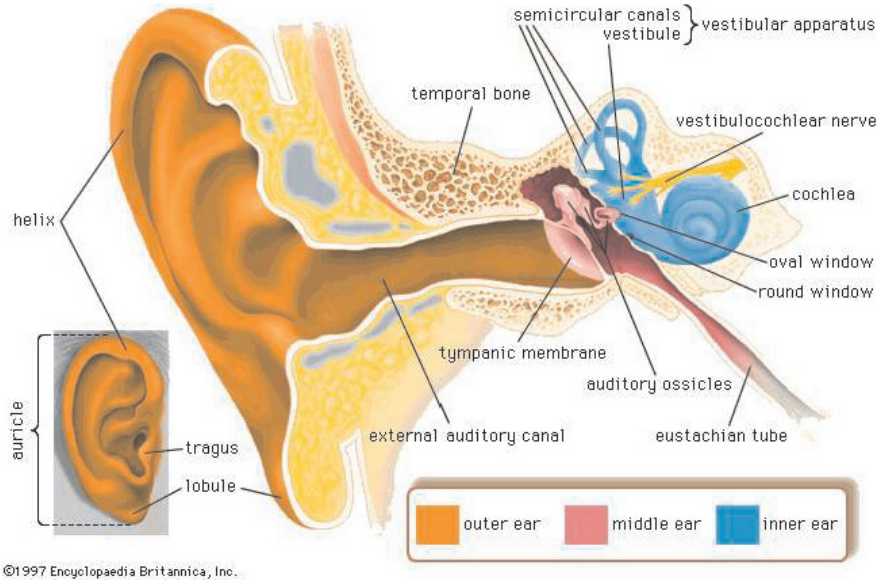

Fig. 2. Human ear (taken with permission from Encyclopaedia Britannica 2001).

Humans have a similar inertial sensing system which is called the *vestibular system* (Gillingham and Previc 1996). Protected inside the bony labyrinth of the temporal bone within the inner ear it has three main parts: the cochlea, the vestibule, and the semicircular canals, see Figure 2. The vestibule houses two otoliths organs, the utricle and the saccule which measure gravitational and inertial forces providing information about the angular position (tilt) and linear motion of the head. The semicircular canals detect angular velocity of the head and are are oriented in three orthogonal planes, thus measuring angular velocity in space.

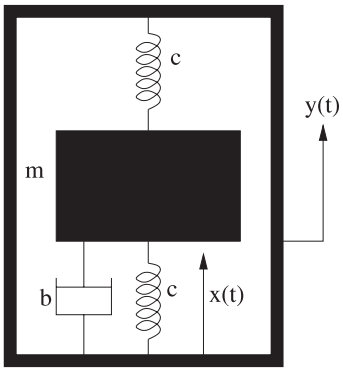

Fig. 3. Principle of single-axis accelerometer.

2.1. Translational Motion

One component of an inertial system is the accelerometer sensor and the basic physical principle, see Figure 3, is quite simple. A proof or seismic mass, m, is supported by an elastic element of stiffness c. This may be a pre-stressed spring or a cantilever beam. A viscous damper, b, provides damping proportional to the relative velocity of the proof mass and the sensor body. The dynamics of this system can be expressed as

$$\ddot{x}(t) + 2\zeta\omega_n\dot{x}(t) + \omega_n^2 x(t0 = -\ddot{y}(t) \qquad (1)$$

which converts acceleration of the sensor body, $\ddot{y}(t)$, to displacement $x(t)$ with a natural frequency ω_n and a damping ratio ζ. Typically the parameters m, c and b are selected to place the resonance well above the motion frequency range of interest. According to the Equivalence principle in general relativity, the effects of gravity and acceleration are the same, that is, we can not determine if the sensor is subject to some component of the gravity vector, or if it is accelerating. Other sensory inputs or strong assumptions are required to resolve this ambiguity.

In engineered systems improvement in surface and bulk micro-machining fabrication methods, along with integrated electronics, have led to the development of low-cost 1, 2 or 3-axis single-chip inertial sensors for applications such as vehicle security, sports training devices, digital camera orientation or laptop drop detection. There are presently three main types of micro-machined low-cost accelerometers: capacitive, piezoelectric and piezo-resistive. The piezoelectric sensors have a large dynamic range but no DC response, making them unsuitable for inertial navigation systems. In the piezo-resistive sensors the position of the proof mass is measured by a piezo-resistor which changes its value. In a capacitive sensor the

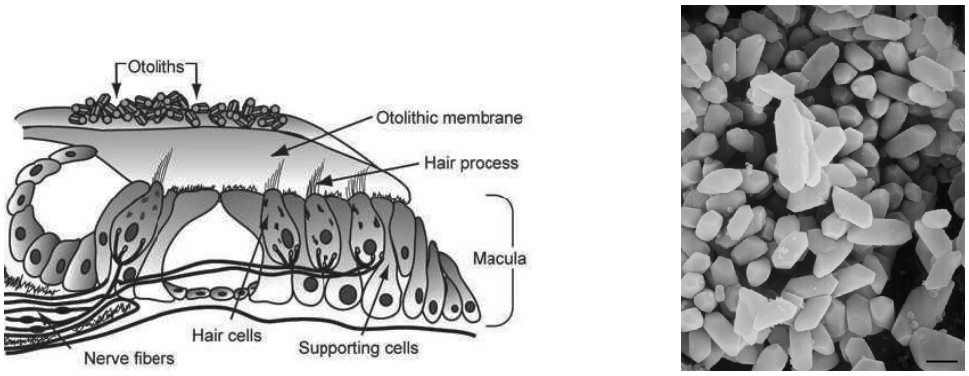

Fig. 4. The accelerometer in the human ear: (left) the overall structure, (right) SEM of gray squirrel otoconia provided by Anna Lysakowski and Steven Price. The bar at bottom right represents 5μm. Originally used as a cover illustration for "Otolith Function in Spatial Orientation and Movement", Vol. 871, Annals of the New York Academy of Sciences, 1999.

proof mass position is determined by changing capacitance. Piezo-resistive sensors require bulk micro-machining, but capacitive sensors can be surface micro-machined providing lower cost sensors will full signal conditioning electronics. A more detailed overview of micro-machined inertial sensors is provided in Yazdi et al. (1998) and Lobo (2002) and the trends in inertial sensors are discussed in Barbour and Schmidt (1999, 2001).

For animals, the accelerometer sensor is remarkably similar to the engineered accelerometer, see Figure 4. The *otolith organs* contain otoliths, literally "ear stones", which are calcium carbonate crystals that serve as the proof mass (Gillingham and Previc 1996). They sit on a gelatinous substance which acts as the spring and damper and in which are embedded hair cells that detect displacement. There are two sensors per ear located inside the semi-circular canal complex, see Figure 6. The *utricle* measures acceleration in the horizontal (front-back) direction, and the *saccule* measures in the vertical direction. To resolve the ambiguity between gravity and body motion, biological systems use other cues such as vision. They also seem to separate the acceleration signals by frequency – the low-frequency component is related to pose, and the high-frequency component is due to acceleration. This evolved assumption is justified since natural motions, such as walking or running, result in acceleration that is typically zero mean over an interval.

2.2. Rotational Motion

When a particle moves in a rotating reference frame, it will experience a Coriolis force

$$\mathbf{F} = 2m\boldsymbol{\omega} \times \mathbf{v}$$

proportional to the velocity $\mathbf{v}$ of the moving particle, the rotation rate of the rotating reference frame $\boldsymbol{\omega}$ and the particle's mass m.

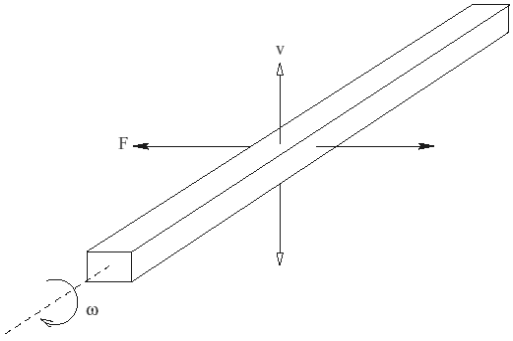

Fig. 5. Principal of vibrating structure gyroscope.

A commonly known device is the gyroscope, essentially a spinning wheel on an axle, invented and named in 1852 by Léon Foucault. The device, once spinning, tends to resist changes to its orientation. For this device the net force is due to all of the particles comprising the spinning disk which yields a moment due to Coriolis acceleration of

$$\mathbf{M} = I\boldsymbol{\omega} \times \mathbf{p}$$

where I is the moment of inertia of the gyroscope disk, $\mathbf{p}$ the angular velocity of the disk and $\boldsymbol{\omega}$ the angular velocity of the gyroscope. As the gyroscope is rotated it exerts an opposing moment which in a strap-down gyroscope configuration is measured, since this is proportional to the angular rate.

The vibrating structure Gyroscope (VSG) has no spinning disk, and is based on producing radial linear motion and measuring the Coriolis effect induced by rotation. Figure 5 represents a VSG where a flexing beam is made to vibrate in the vertical plane. Rotation about the axis of the beam induces a Coriolis force that displaces the beam sideways, which can be detected. MEMS VSG gyroscopes with integrated signal processing electronics in a single piece of silicon are now widely available.

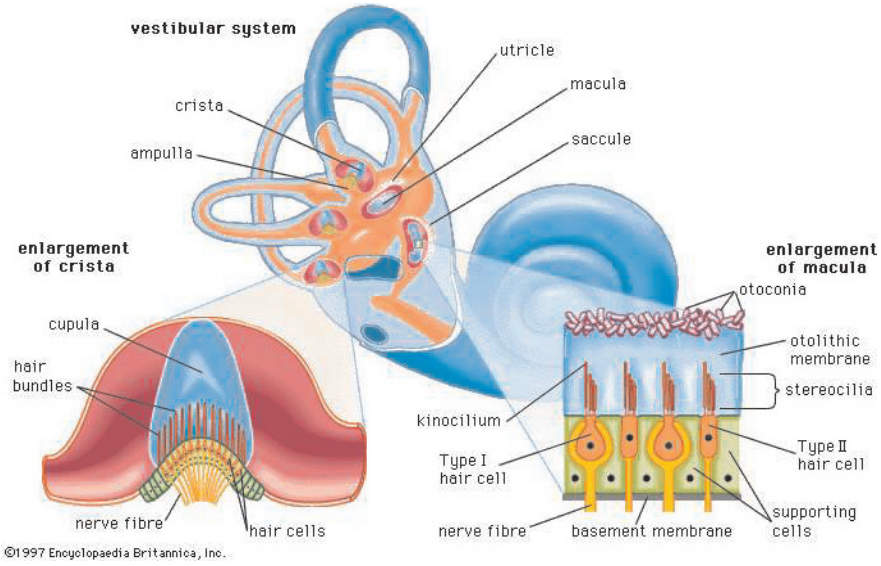

Fig. 6. Human Vestibular System (taken with permission from Encyclopaedia Britannica 2001).

Insects have evolved a similar device — *halteres* are small knobbed structures found as a pair in some two-winged insects (Dickinson 1999). The halteres play an important role in stabilising the gaze of insects during flight and also provide rapid feedback to wing-steering muscles to stabilise aerodynamic force moments.

Vertebrate animals have evolved a sensor based on somewhat different principles and which measures rotational acceleration. Within the labyrinth structure of the ear is the vestibular apparatus which comprises three semicircular canals along with the otolith organs mentioned above (Gillingham and Previc 1996). Each canal is a circular duct filled with a viscous fluid. Rotation causes the fluid to push against one or other end of the duct, where the *ampulla* is located which senses the resulting force, see Figure 6.

Although the semicircular canals are stimulated by angular acceleration, the neural output from the sensory cells in the ampulla represents the velocity at which the canal is being rotated over the range of normal head movements – the canal mechanism performs a mathematical integration of the input signal. This comes about due to the very small internal diameter of the canal, approximately 0.3 mm, which results in a large increase in the viscous properties of the fluid causing *cupula* deflection to be in phase with angular velocity.

Each human ear contains three ducts arranged roughly at right angles with each other so that they represent all three planes in three-dimensional space. The horizontal duct lies in a plane pitched up approximately 30 degrees from the horizontal plane of the head. The anterior canals are located in vertical planes that project forward and outward by approximately 45 degrees, see Figure 7.

The brain combines signals from all six ducts to create a representation of the vector that describes the instantaneous angular velocity of the head. This sensor signal has many function but an important one is to provide a feed-forward signal to the eye muscles to ensure gaze stability, a reflex known as vestibulo-ocular reflex (VOR) that involves two pathways, one direct from the vestibular system to the eye muscles and one via the cerebellum which allows for some measure of gain control (Carpenter 1988).

Human inertial perceptual thresholds are affected by many factors including mental concentration, fatigue, attention and person-to-person variability (Gillingham and Previc 1996). 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 representative perception thresholds for linear acceleration. These are valid for sustained and relatively low frequency stimulus. The currently available low cost inertial sensors are capable of similar performances (Lobo 2002).

2.3. Inertial Navigation

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

$$\mathbf{p} = \int \dot{\mathbf{p}} dt = \iint \ddot{\mathbf{p}} dt = \iint \mathbf{a} dt \qquad (2)$$

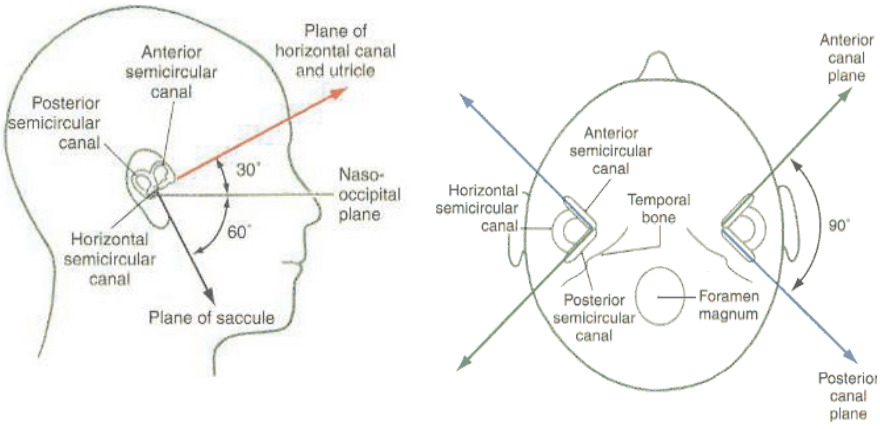

Fig. 7. Axes of the semicircular canals (taken with permission from Dickman 2006).

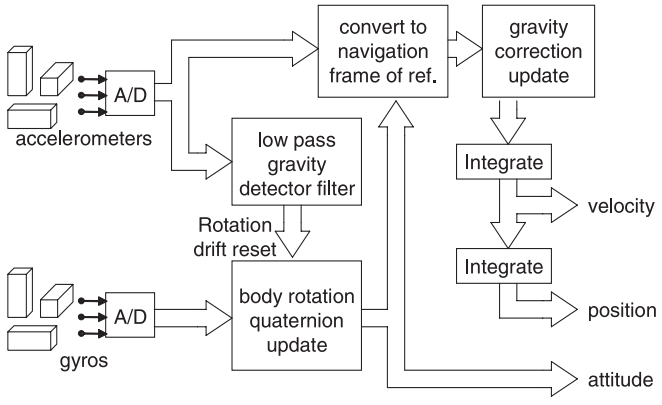

Fig. 8. Simplified strap-down inertial navigation system.

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

The measured accelerations are given in the body frame of reference, initially aligned with the navigation frame of reference. If body rotations occur, they must be taken into account. In gimbaled systems the accelerometers are kept in alignment with the navigation frame of reference and the gyros stabilize the accelerometer platform directly or via a servo system. In a strap-down system the gyros measure the body rotation rate, and the sensed accelerations are computationally converted to the navigation frame of reference. Figure 8 shows a block diagram of a strap-down inertial navigation system. Common complications include variability in the sensor gain and offset, often as a function of temperature, and also cross-axis sensitivity.

The dynamics of our moving sensor system are given by

$$\mathbf{x}_{t+1} = \Phi\mathbf{x}_t + N(0, Q) \tag{3}$$

$$\mathbf{y}_t^i = H\mathbf{x}_t + N(0, R) \tag{4}$$

where the state vector $\mathbf{x}$ comprises system pose and its derivatives in the navigation frame of reference. The observation $\mathbf{y}^i = [\mathbf{a}, \boldsymbol{\omega}]$ are the outputs of the inertial sensors, body acceleration and angular velocity. The angular velocity is integrated to update the rotational attitude of the IMU. Using this attitude, gravity can be computationally separated from the sensed acceleration to yield acceleration of the body itself. Savage describes a complete mechanization using quaternions for performing the rotation update (Savage 1984).

Quaternions provide a convenient representation for 3D rotations (Kuipers 1999). A quaternion $\mathring{\mathbf{q}}$ can be written as

$$\mathring{\mathbf{q}} = q_0 + q_1\mathbf{i} + q_2\mathbf{j} + q_3\mathbf{k} = (q_0, \mathbf{q}) \tag{5}$$

where q_1, q_2 and q_3 are the components of the imaginary or vector part $\mathbf{q}$ of the quaternion, $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are quaternion vector operators, analogous to unit vectors along orthogonal coordinate axes, and q_0 is the scalar part. The quaternion vector operators, which correspond to the $\mathbf{i}$ in complex numbers, are all square roots of -1, and $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$. The magnitude of a quaternion is defined as

$$\|\mathring{\mathbf{q}}\| = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2}. \tag{6}$$

The complex conjugate $\mathring{\mathbf{q}}^*$ of quaternion $\mathring{\mathbf{q}}$ is given by

$$\mathring{\mathbf{q}}^* = q_0 - q_1\mathbf{i} - q_2\mathbf{j} - q_3\mathbf{k} = (q_0, -\mathbf{q}) \tag{7}$$

and the inverse $\mathring{\mathbf{q}}^{-1}$

$$\mathring{\mathbf{q}}^{-1} = \frac{1}{\mathring{\mathbf{q}}} = \frac{\mathring{\mathbf{q}}^*}{\mathring{\mathbf{q}}\mathring{\mathbf{q}}^*}. \tag{8}$$

Vectors can by represented by purely imaginary quaternions. A point in space given by the vector $\mathbf{p}$ can be represented by the quaternion $\mathring{\mathbf{p}} = (0, \mathbf{p})$. In our notation, when multiplying

vectors with quaternions, the corresponding imaginary quaternion is assumed.

Unit quaternions are such that $\|\mathring{\mathbf{q}}\| = 1$ and $\mathring{\mathbf{q}}\mathring{\mathbf{q}}^* = 1$ and for which the inverse is the conjugate $\mathring{\mathbf{q}}^{-1} = \mathring{\mathbf{q}}^*$. Unit quaternions can be used to represent rotations, and the rotation ϕ about a unit vector $\mathbf{u}$ is given by the unit quaternion

$$\mathring{\mathbf{q}} = \cos\frac{\phi}{2} + \sin\frac{\phi}{2}\mathbf{u}. \tag{9}$$

The rotation for a space point, or vector, $\mathbf{p}$ is given by

$$\mathbf{p}' = \mathring{\mathbf{q}}\mathbf{p}\mathring{\mathbf{q}}^{-1} = \mathring{\mathbf{q}}\mathbf{p}\mathring{\mathbf{q}}^*. \tag{10}$$

If the quaternion $\mathring{\mathbf{q}}(k)$ represents the body rotation relative to the navigation frame at sample interval k, then the body accelerations can by converted to the navigation frame of reference by

$$\mathbf{a}_{nav} = \mathring{\mathbf{q}}(k)\,\mathbf{a}_b\mathring{\mathbf{q}}(k)^* \tag{11}$$

In an INS the set of orthogonal gyros provide a measurement of the body rotation rate vector given by

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_x & \omega_y & \omega_z \end{bmatrix}^\mathsf{T} \tag{12}$$

and $\omega = \|\boldsymbol{\omega}\| = \sqrt{\omega_x^2 + \omega_y^2 + \omega_z^2}$ gives the magnitude of the rotation rate and $\frac{\boldsymbol{\omega}}{\omega}$ the unit vector around which the rotation occurs. The rotation increment during a sampling interval Δt is given by the quaternion

$$\Delta\mathring{\mathbf{q}} = \cos\left(\frac{\omega\Delta t}{2}\right) - \sin\left(\frac{\omega\Delta t}{2}\right)\frac{\boldsymbol{\omega}}{\omega} \tag{13}$$

provided that $\omega \neq 0$. The quaternion $\mathring{\mathbf{q}}(k)$, that represents the body rotation relative to the navigation frame at sample interval k, can now be updated by

$$\mathring{\mathbf{q}}(k+1) = \mathring{\mathbf{q}}(k)\,\Delta\mathring{\mathbf{q}} \tag{14}$$

and using (11) the measured body accelerations are converted to the navigation frame, the gravity component is removed, and integration (2) provides body velocity and position in the navigation frame. Typically a unit-quaternion is used and the result of (14) is renormalized to unity to counter numerical effects after each time step.

Referring back to (3) and (4) the state vector is $\mathbf{x} = [\mathbf{p}, \dot{\mathbf{p}}, \ddot{\mathbf{p}}, \mathring{\mathbf{q}}, \dot{\mathring{\mathbf{q}}}]$ where $\mathbf{p} \in \mathbb{R}^3$ and $\mathring{\mathbf{q}} \in \mathrm{SO}(3)$. This can be considered as a state estimation problem, given a dynamic model (3) and the observations (4), and is typically solved using an extended Kalman filter (Jazwinsky 1970). The state vector may be extended to include sensor bias and scale factors.

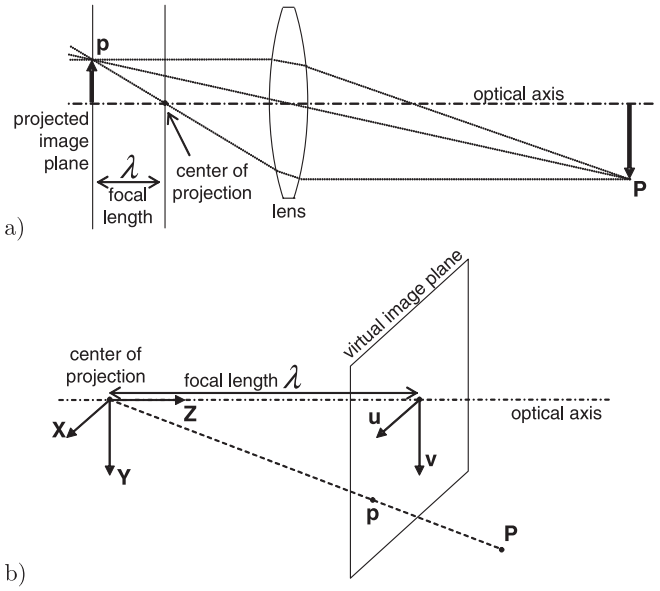

Fig. 9. Camera perspective projection: (a) lens image formation; (b) pinhole camera model.

3. Visual Sensing

3.1. Model of Visual Sensing

The physical principle for image formation on biological and engineered visual sensors is shown in Figure 9. Different geometries can also be considered for the imaging model, such as perspective, fisheye, spherical, catadioptric, etc. (Hartley and Zisserman 2004; Faugeras 1993). Commonly a convex lens projects a 2-dimensional image of the world onto the image plane. The commonly used pinhole camera model, shown in Figure 9, considers one centre of projection where all rays originating 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.

In the pinhole camera model a projection point $\mathbf{p}_i = (u, v)^\mathsf{T}$ in the camera image is related with a 3D point $\mathbf{P} = [X, Y, Z]^\mathsf{T}$ by the perspective relations

$$u = S_u\lambda\frac{X}{Z} + u_0 \quad v = S_v\lambda\frac{Y}{Z} + v_0 \tag{15}$$

where u and v are the picture elements (*pixels*) coordinates on the image plane, $(u_0, v_0)^\mathsf{T}$ is the image center, λ, S_u is the camera focal distance, S_v are the scale factors associated with the physical dimensions of the pixels, and $\mathbf{P}$ is expressed in the camera frame of reference. This camera model ignores lens distortion and assumes there is no skew.

We can rewrite the above equation as

$$u = f_u\frac{X}{Z} + u_0 \quad v = f_v\frac{Y}{Z} + v_0 \tag{16}$$

where u and v are the pixel coordinates with origin at the image center and f is the camera effective focal distance (i.e. includes the pixel scale factor). This can be written as a projective mapping, up to scale factor s as

$$
\begin{aligned}
s\mathbf{p}_i &= \begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \mathbf{CP} \\
&= \underbrace{\begin{bmatrix} f_u & 0 & 0 \\ 0 & f_v & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & u_0 & 0 \\ 0 & 1 & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\text{intrinsic}} \\
&\times \underbrace{[\mathbf{R}\ \mathbf{t}]}_{\text{extrinsic}} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}
\end{aligned} \tag{17}
$$

where $P = (X, Y, Z)$ is the world point, and matrix $\mathbf{C}$ is the projection matrix comprising intrinsic parameters of the camera and lens and the extrinsic parameters of the camera's pose $[\mathbf{R}, \mathbf{t}]$ with respect to the world. The scale factor is arbitrary, and reflects the fact that only the projective ray for each image point is known.

An engineered camera system typically uses a glass lens with fixed focus or some means to move it slightly along the optical axis to achieve a clear image. The image is projected onto a silicon fabricated CMOS or CCD array, with upto 8 million photosites but for low-end cameras is typically only 0.5 million (for VGA resolution). The photosites, of typical dimension $2 - 10\mu$m are arranged in a regular rectangular grid. Low-cost color sensors have color filters printed on the surface, typically filtering a 2x2 group of pixels with one red, one blue and two green filters. Digital processing provides an estimate of red, green and blue intensity at each pixel, but in fact the measurements are not independent. For low-cost sensors exposure control is achieved by altering the integration time and adding an analog gain stage. The dynamic range within an image is typically only 8–10 bits.

The human eye is approximately spherical with a diameter of 15 mm and filled with a clear gel-like material. The lens is a clear, bi-convex structure about 10 mm in diameter, which is deformed by muscles to focus the image which is projected on the retina (image-plane) at the back of the eye. There are two types of photoreceptors in the retina: cones and rods. In normal daylight conditions cone photoreceptors are active and these are color sensitive: 65% sense red, 33% sense green and 2% sense blue. The cones are approximately 3 μm in diameter and 34000 of them are packed into the foveal area of the retina which is only 0.6 mm in diameter. The photoreceptor density in the rest of the retina is considerably lower. The eye has high resolution only over the foveal field of view of a few degrees but subconscious eye motion directs the fovea over the entire field of view. Cone photoreceptors have a dynamic range of 600. At very low light levels the rod photoreceptors become active and provide another factor of 20 in dynamic range. The rod sensors are monochromatic and their density in the fovea is only 7% of that of the cones, but increases in the peripheral region. Their sensitivity is chemically adapted slowly over time. Illumination levels on the retina are controlled by the iris, a muscle-drive diaphram that controls the size of the opening called the pupil. The overall dynamic range of the eye is over 100000 which corresponds to more than 16 bits.

As mentioned earlier, distance information is lost in the projective imaging transformation. Cutting (1997) discusses nine visual cues used by humans to perceive distance, and each cue has a significance that varies with distance. Some cues give ordinal information such as which object is closer than another, whereas others can give more quantitative information. Stereo disparity (Faugeras 1993) is a well-known method to robotics and computer vision researchers to recover distance but it is just one of many perceptual cues used by humans to infer distance and is only effective up to 10 m. The evolutionary development of nine sources of information reflects the importance of depth perception which cannot be trusted to any one single cue.

3.2. Visual Motion

Suppose that the camera is moving with angular velocity $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]$ and translational velocity $\mathbf{T} = [T_x, T_y, T_z]$ with respect to the fixed frame and $\mathbf{P}$ is a point in the world. The velocity of the point $\mathbf{P}$, expressed relative to the camera frame, is given by

$$\dot{\mathbf{P}} = \boldsymbol{\omega} \times \mathbf{P} + \mathbf{T}. \tag{18}$$

Following the approach of Hutchinson et al. (1996) we substitute the perspective projection equations (16) into (18) allowing us to write the derivatives of the coordinates of $\mathbf{P}$ in terms of the image feature parameters u, v as

$$\dot{X} = z\omega_y - \frac{vz}{\lambda}\omega_z + T_x \tag{19}$$

$$\dot{Y} = \frac{uz}{\lambda}\omega_z - z\omega_x + T_y \tag{20}$$

$$\dot{Z} = \frac{z}{\lambda}(v\omega_x - u\omega_y) + T_z. \tag{21}$$

where $f = f_u = f_v$. Our visual feature is the image plane coordinate $[u, v]$ corresponding to $\mathbf{P}$, and using the quotient rule we obtain

$$\dot{u} = f\frac{Z\dot{X} - X\dot{Z}}{Z^2} \quad (22)$$

$$= \frac{f}{Z^2}\left\{Z\left[Z\omega_y - \frac{vz}{f}\omega_z + T_x\right]\right.$$

$$\left. - \frac{uz}{f}\left[\frac{Z}{f}(v\omega_x - u\omega_y) + T_z\right]\right\} \quad (23)$$

$$= \frac{f}{Z}T_x - \frac{u}{Z}T_z - \frac{uv}{f}\omega_x + \frac{f^2+u^2}{f}\omega_y - v\omega_z \quad (24)$$

and similarly

$$\dot{v} = \frac{f}{Z}T_y - \frac{v}{Z}T_z + \frac{-f^2-v^2}{f}\omega_x + \frac{uv}{f}\omega_y + u\omega_z. \quad (25)$$

Rewriting these two equations in matrix form we obtain

$$\begin{bmatrix}\dot{u}\\ \dot{v}\end{bmatrix} = \begin{bmatrix} \frac{f_x}{z} & 0 & \frac{-u}{z} & \frac{-uv}{f_x} & \frac{-\left(f_x^2+u^2\right)}{f_x} & -v \\ 0 & \frac{f_y}{z} & \frac{-v}{z} & \frac{\left(f_y^2+u^2\right)}{f_y} & \frac{uv}{f_y} & u \end{bmatrix} \times \begin{bmatrix} T_x \\ T_y \\ T_z \\ \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} \quad (26)$$

which relates image-plane velocity of a point to the relative velocity of the point with respect to the camera through the image Jacobian matrix. We can clearly see that image-plane velocity is the summation of 6 motion components, making it impossible to disambiguate rotational from translational motion from the observed motion of a single point. Also we can see that the first three columns in the image Jacobian have the range in the denominator. This represents the effect of range, z, on apparent velocity, i.e. a slow close object appears to move at the same speed as a distant fast object.

3.3. Concurrent Estimation of Structure and Motion

While direct depth information is lost from a single perspective view, multiple views from different viewpoints hold the possibility of recovering depth, and this is the well-studied computer vision problem called structure from motion (SFM) (Jebara et al. 1999; Huang and Netraveli 1994; Hartley and Zisserman 2004; Ma et al. 2004). More formally, SFM estimates both the 3D position of points in the scene with respect to some fixed coordinate frame and also the pose of the camera for each frame. A typical SFM implementation has the following components:

1. Robust detection of salient features, such as points or lines, in each scene that we can observe across multiple consecutive images.
2. Determining the correspondence of these features between consecutive images.
3. Updating the estimate of scene structure and camera pose.

Within this general approach many variations are reported in the literature.

Point features are most commonly used and the Kanade–Lucas tracker (Kanade and Lucas 1981) combines corner detection with tracking, yielding multi-frame tracks of features on the image plane in an efficient way for small motions that typically occur between consecutive video frames. Alternatively corner features can be found using the Shi–Tomasi (Tomasi and Shi 1994) or Harris (Harris and Stephens 1988) detectors in individual frames and then correspondence is determined, generally involving an exhaustive comparison of all features between consecutive image frames (Ma et al. 2004; Nistér et al. 2006). Point feature comparison is typically based on the similarity of regions surrounding each corner. The search process can be pruned by assuming the image-plane motion is small. Information about camera motion from the inertial sensors can also be used to predict feature position, which can significantly reduce the search space, see for example Corke (2004). Methods have been proposed that do not use correspondence such as Dellaert (2000), or which use a probabilistic measure of correspondence such as Domke and Aloimonos (2005, 2006). Robust feature detection is discussed further in Vincze and Hager 1999), and implementations of trackers are generally available (Hager and Toyama 1998; Xvision2; Birchfield). SFM using lines is discussed in Rehbinder and Ghosh (2003) and Huang and Netraveli (1994).

Assuming a rigid scene, a small number of corresponding points from 2 or 3 consecutive images can be used to estimate the change in camera pose and the world coordinates of the points (Nistér et al. 2006; Ma et al. 2004). Techniques such as random sample consensus (RANSAC) (Fischler and Bolles 1981) or least-median squares (Zhang et al. 1995) are applied to provide robustness against correspondence errors. Nistér presents a pre-emptive technique which gives high efficiency in limited time as required for real-time applications (Nistér 2005). A subsequent smoothing filter can be applied to the camera motion to account for dynamic constraints (Soatto et al. 1993). Alternatively the problem can be formulated as estimating the state of the dynamic system

$$\mathbf{x}_{t+1} = \Phi\mathbf{x}_t + N(0, Q) \quad (27)$$

$$\mathbf{y}_t^v = H(\mathbf{x}_t) + N(0, R) \quad (28)$$

where the state vector $\mathbf{x} = [\mathbf{x^c}|\mathbf{x^w}]$ comprises the state of the camera $\mathbf{x}^c = [\mathbf{P}, \mathring{\mathbf{q}}]$ and the state of the world

$\mathbf{x}^w = [\mathbf{P}_1, \mathbf{P}_2, \cdots \mathbf{P}_N]$, the 3D coordinates of N scene points. The camera state includes camera pose, and if the camera is uncalibrated or partly calibrated will also include the unknown intrinsic parameters. Many representations of the world coordinates have been proposed, including Cartesian (x, y, z), image-plane coordinate augmented by depth along the ray (u, v, d) (Jebara et al., 1999), and a probabilistic depth distribution (Davison 2003). The observation $\mathbf{y}_t^v = [(u_1, v_1), (u_2, v_2), \cdots (u_N, v_N)]$ represents the image-plane coordinates of the world points, which is typically a very non-linear function of camera pose and depends on the type of camera projection model. Most SFM systems use conventional projective cameras but wide angle lenses have been used by Davison et al. (2004) and Strelow (2004). Various approaches to solving this estimation problem have been demonstrated, including extended Kalman filters by Broida et al. (1990) and Azarbayejani and Pentland (1995). A software toolkit for implementing SFM is available (Torr).

There are very strong similarities between SFM and the simultaneous localization and mapping (SLAM) problem (Newman 2007; Thrun et al. 2005; Montemerlo and Thrun 2007), also known as concurrent mapping and localization (CML). A SLAM algorithm incrementally builds a stochastic map, with every new data acquisition it estimates the robot pose from the matches between observed and previously perceived landmarks, and updates the map with new landmarks and fused estimates for matched ones.

With a single camera there is no depth data for the landmark, and this is essentially the bearing-only SLAM problem. To determine the initial estimate of landmark depth, a landmark initialization process is used which combines at least two observations of the same features from far enough apart robot poses. Real-time visual SLAM has recently been accomplished using a single camera (Davison 2003; Davison et al. 2004, 2007) or with a stereo camera (Molton and Brady 2000) which provides range and bearing observations. Many different sensors can be used to detect landmarks, but to apply the classic extended Kalman filter (EKF) SLAM algorithm, the landmark addition into the stochastic map requires a full Gaussian estimation of its state.

4. Inertial and Visual Sensor Fusion

Combining camera and inertial sensors exploits their complementary characteristics:

1. The inertial sensor is unable to distinguish a change in inclination from acceleration of the body, due to Einstein's equivalence principle.

2. Inertial sensors have large measurement uncertainty at slow motion and lower relative uncertainty at high velocities. Inertial sensors can measure very high velocities and accelerations.

3. Cameras can track features very accurately at low velocities. With increasing velocity, tracking is less accurate due to motion blur and the effect of camera sampling rate. For high velocities and accelerations cameras with higher frame rate can be used up to a limit, but the increase in bandwidth complicates real-time implementations.

4. In a projective image we cannot, according to (26), disambiguate rotational from translational motion.

5. A near object with low relative speed appears the same as a far object with high relative speed, again according to (26).

Thus the motivation for integration of vision and inertial sensing is clear. Starting with the early work of researchers such as Viéville and Faugeras (1990) there is now growing interest and application of inertial and visual fusion which is driven by the availability of small and low-cost sensors.

4.1. Gravity as a Vertical Reference

In vision based systems used in mobile robotics, the perception of self-motion and structure of the environment is essential. Inertial sensors can provide valuable data not only about camera ego-motion, but also an absolute reference for how to expect features and structures to be oriented in the world.

A static camera is capable of observing one important inertial cue – gravity. The vertical vanishing point of any vertical world features defines the gravity reference for the camera. The image horizon line is another cue for camera attitude. The path of objects in free fall or ballistic motion also provide a vertical reference. With some prior knowledge about expected scene features, the visual gravity cues can be detected and a vertical reference defined for the camera. Conversely, having the vertical reference from static inertial sensors provides knowledge about expected scene features. Vision processing can use this external reference for feature extraction, simplifying correspondence, object identification and scene interpretation.

In dynamic systems keeping track of the vertical direction is required, so that gravity acceleration can be compensated for, and it also provides a valuable spatial reference. Dynamic inertial cues also provide an image independent location of the image focus of expansion and center of rotation which can be useful during visual based navigation tasks.

Low level monocular image processing can use the vertical reference to tune edge detection to find relevant features such as vertical or horizontal scene elements. In stereo vision the vertical reference provides an external restriction when considering ground plane or levelled plane point correspondence in the stereo pair. Results for ground plane segmentation of

feature points, vertical line detection and 3D vertical line segmentation are presented in Lobo and Dias (2003).

In Lobo and Dias (2004) depth maps obtained from stereo vision are rotated to a vertically aligned world frame of reference using the inertial vertical reference. Segmentation of planar levelled patches is simplified, and 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. In Viéville et al. (1995) ego motion is estimated using the vertical cue. Using the vertical as a basic cue for 3D orientation simplifies the structure from motion paradigm. A line segment based module to recover ego motion is implemented that concurrently builds a 3D map of the environment in which the absolute vertical is taken into account. The proposed method reduces the disparity between two frames in such a way that 3D vision is simplified. In particular the correspondence problem is simplified.

Gravity provides a valuable spatial reference, however for rotations about a vertical axis gravity provides no cues, and gyro integration is required to keep track of body attitude. The earth's magnetic field can be used to provide the missing bearing (Caruso et al. 1998), but magnetic sensing is sensitive to nearby ferrous metals and electric currents. In fact, there is some overlap and complementarity between the two sensors, with different noise characteristics that can be exploited to provide a useful rotation absolute reference as in Roetenberg et al. (2003, 2005).

4.2. Concurrent Estimation of Structure and Motion

According to Qian et al. (2001) the advantages of inertial and visual fusion in SFM are: greater robustness to feature tracking errors, fewer features required to recover camera motion and reduced ambiguity in the recovery of camera motion. There are two broad approaches that we will call *loosely* and *tightly* coupled. The loosely coupled approach uses separate INS and SFM blocks, running at different rates and exchanging information. The tightly-coupled systems combine the disparate raw data of vision and inertial sensors in a single, optimum filter, rather than cascading two filters, one for each sensor.

4.2.1. Loosely Coupled Systems

In the loosely coupled approach, see Figure 10, the INS and SFM blocks run independently. Translational and angular velocity estimates from the INS are used to predict feature motion, and velocity estimates from SFM can be used to bound integration errors in the INS. Prediction of feature motion provides a virtual stabilized camera, which has the advantages of low-cost, small-size, no moving parts and superiour dynamics compared to a mechanical pan/tilt camera. This makes the feature correspondence process more robust and can reduce the search space thus reducing computational load.

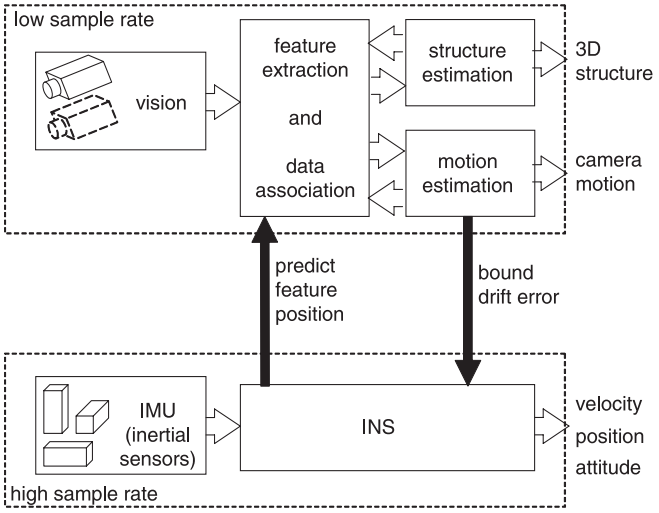

Fig. 10. Loosely coupled inertial-visual system.

Primates however do have the equivalent of an active pan/tilt camera system. The vestibulo-ocular reflex (Carpenter 1988) provides a feedforward from head rotational velocity (sensed in the semi-circular canals) to eye rotational velocity. A simple demonstration shows the effectiveness of VOR for retinal-image stabilization. Hold your extended fingers at arms length in front of your face, and move them slowly from side to side. You can clearly see them because of your visual (optokinetic) tracking reflexes. However as the frequency of movement increases you will reaches a point where the fingers cannot be seen clearly – they are blurred by the movement – typically around 60 deg.s^{-1} or 1 or 2Hz for most people. Now, if the fingers are held still and the head is rotated back and forth at that frequency the fingers remain perfectly clear – this is VOR in action.

Conflicts between these two subsystems, visual and vestibular, lead to interesting physiological effects. The sensation of vertigo, when looking down from a high place, occurs as the body tends to sway in order to obtain a visual stimulus since the viewed scene is very far away. Even large amplitude movements fail to provide any visual stimulus, but the large swaying motion triggers the vestibular system, giving an alarm that the body is out of balance. During prolonged head rotation the elasticity of the cupula gradually returns it to its resting position, signaling no rotation. This conflicts with information from the eyes and causes the sensation of dizziness. Motion sickness results from conflicts between these sensors, typically when the vestibular system indicates motion, but the eyes do not.

For low frequency motion of external world features relative to the body, or body motion relative to the world, gaze stabilization is done by the visual system with the optokinetic tracking reflexes. As the frequency increases, the

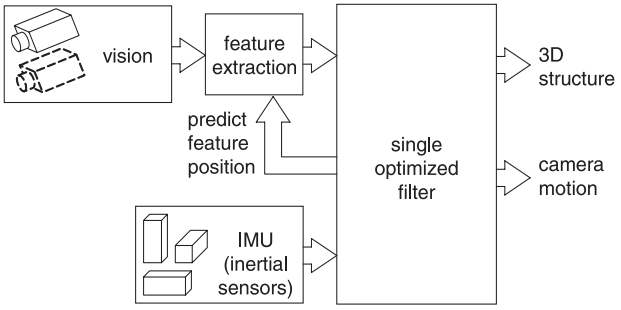

Fig. 11. Tightly coupled inertial-visual system.

vestibulo-ocular reflexes assume a predominant role. In normal human activity the higher frequencies of relative motion are due to head and body motion, where the vestibular system can provide a suitable stimulus for the gaze stabilization reflexes. In engineering terms this is an example of complementary filtering (Zimmerman and Sulzer 1991) to fuse the inertial (rate) and visual (position) data, see for example, Corke (2004).

4.2.2. Tightly Coupled Systems

In the tightly coupled approach, shown in Figure 11, a single high-order estimation filter is used. Combining the observation equations (4) and (28) we can write

$$\mathbf{x}_{t+1} = \Phi \mathbf{x}_t + N(0, Q) \tag{29}$$

$$\mathbf{y}_t^i = H^i(\mathbf{x}_t) + N(0, R) \tag{30}$$

$$\mathbf{y}_t^v = H^v(\mathbf{x}_t) + N(0, R) \tag{31}$$

where the state vector $\mathbf{x} = [\mathbf{x^c}|\mathbf{x^w}]$ comprises the inertial state of the camera $\mathbf{x}^c = [\mathbf{p}, \dot{\mathbf{p}}, \ddot{\mathbf{p}}, \mathring{\mathbf{q}}, \dot{\mathring{\mathbf{q}}}]$ and the state of the world $\mathbf{x}^w = [\mathbf{P}_1, \mathbf{P}_2, \cdots \mathbf{P}_N]$, and the 3D coordinates of N scene points. The inertial observations $\mathbf{y}^i = [\ddot{\mathbf{p}}, \boldsymbol{\omega}]$ are the outputs of the inertial sensors, while the visual observations, $\mathbf{y}^v$ are the image-plane coordinates of the world points. Additionally the state vector may be augmented by unknown camera intrinsic parameters and inertial sensor bias and scale parameters. The state vector will be large, typically over 20 states, which has implications for computational load and for tuning. Implementation is complicated by the fact that the visual and inertial observations occur at quite different rates.

This estimation problem has been solved using offline batch optimzation by Strelow (2004), extended Kalman filter by Qian et al. (2001), iterated extended Kalman filter by Strelow (2004) and unscented Kalman filter by Julier et al. (2000) and Armesto et al. (2004). Strelow (2004) also investigates the use of the panoramic catadioptric camera combined with inertial sensing.

4.3. A Summary of Applications and Related Work

Integration of visual and inertial sensing modalities opens new application directions in robotics and other fields. Inertial sensor technology has been steadily improving (Yazdi et al. 1998; Barbour and Schmidt 2001), enabling innovative applications such as the development of vestibular prostheses for human patients (Wall et al. 2003). This section briefly summarizes various applications of inertial and visual sensors reported in the literature, such as virtual and augmented reality, localization and mapping for navigation, involving gaze control, pose and motion estimation, hybrid trackers and structure from motion.

To better exploit the benefits of combining the two sensing modalities in artificial systems, a clear understanding of biological systems provides useful perspective. Taking advantage of improved brain imaging techniques, a better understanding of the visual motion and self-movement interactions has been pursued (Beer et al. 2002; Previc et al. 2000). Vestibular information is necessary not only for vestibular reflexes but also in various cognitive functions for our adequate behavior in three-dimensional space. In Fukushima (1997) the regions of the cerebral cortex where vestibular information is represented is investigated. Perception and action influence each other, making some biological systems highly coupled and complex, from which direct models for sensor fusion are not easily derived (Hurley 2001). In Leone (1998) and Angelaki et al. (1999) the role of gravity in visual perception and how the brain deals with the ambiguity between inclination and body acceleration is investigated. In Harris et al. (2000) and Reymond et al. (2002) the motion perception inferred from visuovestibular cues is studied. The perceived relative motion is important for posture control (Kelly et al. 2005).

In Viéville and Faugeras (1989) the use of inertial sensors in computer vision applications was proposed, and further work 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 (Viéville and Faugeras 1990; Viéville et al. 1993a, 1993b, 1995; Viéville 1997). In Lobo and Dias (2003) a framework is proposed 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 Lobo et al. (2003) world vertical feature detection and 3D mapping is presented, and in Lobo and Dias (2004) the inertial vertical reference is used to improve depth map alignment and registration.

An important aspect in practical implementation is system calibration. When visual and inertial sensors are integrated in a system their relative pose must be determined. A specific calibration stand with a target board with a set of coded fiducials is used in Foxlin and Naimark (2003a) to fully calibrate a miniaturized hybrid self-tracker system. In Lang and Pinz (2005) the rotation calibration between sensors is based on rotation dif-

ferences. In Lobo and Dias (2005, 2007) a simple relative pose calibration procedure based on observing gravity is proposed, and a calibration toolbox is provided (Lobo 2006).

Some bio-inspired robotic implementations of image stabilization and gaze control have been proposed (Panerai and Sandini 1998; Panerai et al. 2000, 2002; Viollet and Franceschini 2005) that try to mimic the vestibulo-ocular reflex, using inertial sensors to generate compensatory movements and limit the amplitude of image motion to a range can be dealt by the visual algorithms.

Virtual reality applications have always required motion sensors on the user which is inconvenient. Augmented reality, where virtual reality is overlaid onto a realtime view, is particularly sensitive to any mismatch between real and estimated user motion. Precise user attitude and translation can be obtained with several sensor suites, using external vision and specific markers, radio transponders, ultrasound beacons, laser beacons, etc. Aiming for low cost self contained systems, MEMs inertial sensors are being used in combination with computer vision techniques (You et al. 1999). The ultimate goal is to have a visuo-inertial tracker, that can operate in arbitrary unprepared environments relying on natural features, suitable for augmented reality applications. In You and Neumann (2001) a two-channel complementary motion extended Kalman filter is used to combine the low-frequency stability of vision sensors with the high-frequency tracking of gyroscope sensors, hence, achieving stable static and dynamic six-degree-of-freedom pose tracking. Augmented reality systems rely on hybrid trackers to successfully fuse real time imagery with dynamic 3D model (You et al. 1999; Lang et al. 2002; Neumann et al. 2003; Jiang et al. 2004).

Many other hybrid self-trackers based on inertial and vision sensors have been proposed (Hoff et al. 1996; Azuma et al. 1999; Chai et al. 2002; Naimark and Foxlin 2002; Foxlin and Naimark 2003b; Ribo et al. 2004; Hogue et al. 2004; Alenya et al. 2004; Klein and Drummond 2004). The visual tracking relies on either specific targets, line contours or more demanding natural landmarks, and both visual and inertial estimators interact to produce a hybrid tracker. Some commercial hybrid self-tracker systems are being developed such as Foxlin and Naimark (2003b) and Foxlin et al. (2004). In Grimm and Grigat (2004) the pose of an ergonomic pen-like human–computer interface is tracked in real time using vision and a set of accelerometers.

As mentioned above, the integration of inertial sensors can reduce ambiguities and improve robustness of structure from motion methods (Qian et al. 2001, 2002). The dual problem of motion estimation from observed structure has long been pursued, and some recent work that explores the complementarity of inertial and visual sensing for motion estimation is Jung and Taylor (2001), Strelow and Singh (2002, 2003), Chroust and Vincze (2004), and Chen and Pinz (2004).

Applications to robotics are increasing. Some early work on vision systems for automated passenger vehicles also incorporated inertial sensors and explored the benefits of visuo-inertial tracking (Dickmanns 1998; Goldbeck et al. 2000). Other ground vehicle applications include agricultural vehicles (Hague et al. 2000), wheelchairs (Goedem et al. 2004) and indoor mobile robots (Stratmann and Solda 2004; Diel et al. 2005). Recent work related to aerial vehicles (UAVs) includes fixed wing aircraft (Kim and Sukkarich 2004, 2007); Bryson and Sukkarich 2007; Nygards et al. 2004; Graovac 2004), rotorcraft (Muratet et al. 2005; Corke 2004) and descending spacecraft (Roumeliotis et al. 2002). For underwater vehicles (AUVs) there are recent results in Eustice et al. (2005), Huster and Rock (2003) and Dunbabin et al. (2006).

5. Conclusion

This paper has presented a tutorial introduction to inertial and visual sensors and discussed how they may be fused to create a robust estimate of self motion. For mobile robotics, ground, air and underwater, a sense of position (localization) and motion are critically important. Biological systems from flying insects to humans have evolved complementary sensor systems that provide this functionality which is a testament to their utility. Artificial systems should also exploit this sensor fusion. Inertial sensors coupled to cameras can provide valuable data about camera ego-motion and how world features are expected to be oriented. Feature detection and tracking benefits from both static and dynamic inertial information. Visual and inertial sensors today have high performance and are low cost and compact. They require no external reference and emit no radiation. These sensors have useful complementarities, each able to cover the limitations and deficiencies of the other.

References

Alenya, G., Martnez, E., and Torras, C. (2004). Fusing visual and inertial sensing to recover robot ego-motion. *Journal of Robotic Systems*, **21**(1): 23–32.

Angelaki, D. E., McHenry, M. Q., Dickman, J. D., Newlands, S. D., and Hess, B. J. (1999). Computation of inertial motion: neural strategies to resolve ambiguous otolith information. *The Journal Of Neuro-science: The Official Journal Of The Society For Neuroscience*, **19**(1): 316–327.

Armesto, L., Chroust, S., Vincze, M., and Tornero, J. (2004). Multi-rate fusion with vision and inertial sensors. *Proceedings of Iinternational Conference on robotics and Automation*, pp. 193–199.

Azarbayejani, A. and Pentland, A. P. (1995). Recursive estimation of motion, structure, and focal length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(6): 562–575.

Azuma, R., Hoff, B., Neely, H., and Sarfaty, R. (1999). A motion-stabilized outdoor augmented reality system. *Proceedings of IEEE Virtual Reality*, pp. 252–259.

Barbour, N. and Schmidt, G. (1999). Inertial sensor technology trends. *The Draper Technology Digest*, **3**: 5–13.

Barbour, N. and Schmidt, G. (2001). Inertial sensor technology trends. *IEEE Sensors Journal*, **1**(4): 332–339.

Beer, J., Blakemore, C., Previc, F., and Liotti, M. (2002). Areas of the human brain activated by ambient visual motion, indicating three kinds of self-movement. *Experimental Brain Research*, **143**(1): 78–88.

Birchfield, S. Implementation of klt tracker. http://www.ces.clemson.edu/ stb/klt/.

Broida, T., Chandrashekhar, S., and Chellappa, R. (1990). Recursive 3-d motion estimation from a monocular image sequence. *IEEE Transactions on Aerospace and Electronic Systems*, **26**(4): 639–656.

Bryson, M. and Sukkarieh, S. (2007). Building a robust implementation of bearing-only inertial slam for a uav. *Journal of Field Robotics*, **24**(1–2): 113–143.

Carpenter, H. (1988). *Movements of the Eyes*, 2nd edn. Pion Limited, London, ISBN 0-85086-109-8.

Caruso, M. J., Bratland, T., Smith, C. H., and Schneider, R. (1998). A New Perspective on Magnetic Field Sensing. Technical report, Honeywell Inc.

Chai, L., Hoff, W. A., and Vincent, T. (2002). Three-dimensional motion and structure estimation using inertial sensors and computer vision for augmented reality. *Presence*, **11**(5): 474–492.

Chen, J. and Pinz, A. (2004). Structure and motion by fusion of inertial and vision-based tracking. In *Digital Imaging in Media and Education* (eds W. Burger and J. Scharinger), Vol. 179 of Schriften-reihe, pp. 55–62. OCG. Proceedings of the 28th ¨OAGM/AAPR Conference.

Chroust, S. G. and Vincze, M. (2004). Fusion of vision and inertial data for motion and structure estimation. *Journal of Robotic Systems*, **21**(2): 73–83.

Corke, P. (2004). An inertial and visual sensing system for a small autonomous helicopter. *Journal of Robotic Systems*, **21**(2): 43–51.

Cutting, J. (1997). How the eye measures reality and virtual reality. In *Behavior Research Methods, Instrumentation and Computers*, pp. 29–36. Cornell University.

Davison, A., Cid, Y. G., and Kita, N. (2004). Real-time 3D SLAM with wide-angle vision. *Proceedings of IFAC Symposium on Intelligent Autonomous Vehicles*, Lisbon.

Davison, A. J. (2003). Real-time simultaneous localisation and mapping with a single camera. *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, USA, IEEE Computer Society, p.1403.

Davison, A. J., Reid, I., Molton, N., and Stasse, O. (2007). MonoSLAM: real-time single camera SLAM. Pre-print accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence.

Dellaert, F., Seitz, S., Thorpe, C., and Thrun, S. (2000). Structure from motion without correspondence. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'00)*.

Dickinson, M. H. (1999). Haltere-mediated equilibrium reflexes of the fruit fly, drosophila melanogaster. *Philosophical Transactions: Biological Sciences*, **354**(1385): 903–916.

Dickman, J. D. (2006). The vestibular system. In *Fundamental Neuroscience for Basic and Clinical Applications*, 3rd edn, pp. 350–366, Churchill Livingstone/Elsevier, Philadelphia, PA.

Dickmanns, E. D. (1998). Vehicles capable of dynamic vision: a new breed of technical beings? *Artificial Intelligence*, **103**(1–2): 49–76.

Diel, D. D., DeBitetto, P., and Teller, S. (2005). Epipolar constraints for vision-aided inertial navigation. *Proceedings of IEEE Motion and Video Computing*, pp. 221–228.

Domke, J. and Aloimonos, Y. (2005). A probabilistic framework for correspondence and egomotion. *ICCV Workshop on Dynamic Vision*.

Domke, J. and Aloimonos, Y. (2006). A probabilistic notion of correspondence and the epipolar constraint. *3DPVT – International Symposium on 3D Data Processing Visualization and Transmission*.

Dunbabin, M., Usher, K., and Corke, P. (2006). Visual motion estimation for an autonomous underwater reef monitoring robot. *Field and Service Robotics: Result of the 5th International Conference* (eds P. Corke and S. Sukkariah), Vol. 25 of STAR, pp. 31–42. Springer.

Encyclopaedia Britannica (2001). www.britannica.com.

Eustice, R., Singh, H., Leonard, J., Walter, M., and Ballard, R. (2005). Visually navigating the RMS titanic with SLAM information filters. *Proceedings of Robotics: Science and Systems (RSS)*, Cambridge, MA, USA.

Faugeras, O. (1993). *Three-dimensional Computer Vision: a Geometric Viewpoint*. MIT Press, Cambridge, MA.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, **24**(6): 381–395.

Foxlin, E., Altshuler, Y., Naimark, L., and Harrington, M. (2004). Flighttracker: A novel optical/inertial tracker for cockpit enhanced vision. *Proceedings of Third IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 212–221.

Foxlin, E. and Naimark, L. (2003a). Miniaturization, calibration and accuracy evaluation of a hybrid self-tracker. *Proceedings of The Second IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 151–160.

Foxlin, E. and Naimark, L. (2003b). Vis-tracker: A wearable vision-inertial self-tracker. *Proceedings of IEEE Virtual Reality 2003*, IEEE Computer Society, p. 199.

Fukushima, K. (1997). Corticovestibular interactions: anatomy, electrophysiology, and functional considerations. *Experimental Brain Research*, **117**(1): 1–16.

Gillingham, K. K. and Previc, F. H. (1996). Spatial orientation in flight. In *Fundamentals of Aerospace Medicine*, 2nd edn (ed. R. L. DeHart), Chapter 11, Williams and Wilkins.

Goedem, T., Nuttin, M., Tuytelaars, T., and Gool, L. V. (2004). Vision based intelligent wheel chair control: The role of vision and inertial sensing in topological navigation. *Journal of Robotic Systems*, **21**(2): 85–94.

Goldbeck, J., Huertgen, B., Ernst, S., and Kelch, L. (2000). Lane following combining vision and dgps. *Image and Vision Computing*, **18**(5): 425–433.

Graovac, S. (2004). Principles of fusion of inertial navigation and dynamic vision. *Journal of Robotic Systems*, **21**(1): 13–22.

Grimm, M. and Grigat, R.-R. (2004). Real-time hybrid pose estimation from vision and inertial data. *Proceedings of First Canadian Conference on Computer and Robot Vision*, pp. 480–486.

Hager, G. and Toyama, K. (1998). X vision: A portable substrate for real-time vision applications. *Computer Vision and Image Understanding*, **69**(1): 23–37.

Hague, T., Marchant, J. A., and Tillett, N. D. (2000). Ground based sensing systems for autonomous agricultural vehicles. *Computers and Electronics in Agriculture*, **25**(1–2): 11–28.

Harris, C. G. and Stephens, M. J. (1988). A combined corner and edge detector. *Proceedings of the Fourth Alvey Vision Conference*, Manchester, pp. 147–151.

Harris, L. R., Jenkin, M., and Zikovitz, D. C. (2000). Visual and non-visual cues in the perception of linear self motion. *Experimental Brain Research*, **135**(1): 12–21.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*, 2nd edn. Cambridge University Press, ISBN: 0521540518.

Hoff, W. A., Nguyen, K., and Lyon, T. (1996). Computer vision-based registration techniques for augmented reality. *Proceedings of Intelligent Robots and Computer Vision*, pp. 538–548.

Hogue, A., Jenkin, M., and Allison, R. (2004). An opticalinertial tracking system for fully-enclosed vr displays. *Proceedings of the First Canadian Conference on Computer and Robot Vision*, pp. 22–29.

Huang, T. and Netravali, A. (1994). Motion and structure from feature correspondences: A review. *Proceedings of IEEE*, **82**(2): 252–268.

Hughes, H. (1999). *Sensory Exotica: A World Beyond Human Experience*, MIT Press, Cambridge, MA.

Hurley, S. (2001). Perception and action: alternative views. *Synthese*, **129**(1): 3–40.

Huster, A. and Rock, S. (2003). Relative position sensing by fusing monocular vision and inertial rate sensors. *Proceedings of 11th International Conference on Advanced Robotics*, Coimbra, Portugal, pp. 1562–1567.

Hutchinson, S., Hager, G., and Corke, P. (1996). A tutorial on visual servo control. *IEEE Transactions on Robotics and Automation*, **12**(5).

Jazwinsky, A. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

Jebara, T., Azarbayejani, A., and Pentland, A. (1999). 3D structure from 2D motion. *IEEE Signal Processing Magazine*, **16**(3).

Jiang, B., Neumann, U., and You, S. (2004). A robust hybrid tracking system for outdoor augmented reality. *Proceedings of IEEE Virtual Reality*, pp. 3–275.

Julier, S., Uhlmann, J., and Durrant-Whyte, H. (2000). A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control*, **45**(3): 477–482.

Jung, S.-H. and Taylor, C. (2001). Camera trajectory estimation using inertial sensor measurements and structure from motion results. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. II–732–II–737.

Kanade, T. and Lucas, B. (1981). An iterative image registration technique with an application to stereo vision. *IJCAI* 81, pages 674–679.

Kelly, J.W., Loomis, J. M., and Beall, A. C. (2005). The importance of perceived relative motion in the control of posture. *Experimental Brain Research*, **161**(3): 285–292.

Kim, J. and Sukkarieh, S. (2004). Improving the real-time efficiency of inertial slam and understanding its observability. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vol. 1, pp. 21–26.

Kim, J. and Sukkarieh, S. (2007). Real-time implementation of airborne inertial-slam. *Robotics and Autonomous Systems*, **55**(1): 6271.

Klein, G. S. W. and Drummond, T. W. (2004). Tightly integrated sensor fusion for robust visual tracking. *Image and Vision Computing*, **22**(10): 769–776.

Kuipers, J. B. (1999). *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace and Virtual Reality*. Princeton University Press, Princeton, NJ.

Lang, P., Kusej, A., Pinz, A., and Brasseur, G. (2002). Inertial tracking for mobile augmented reality. *Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference*, Vol. 2, pp. 1583–1587.

Lang, P. and Pinz, A. (2005). Calibration of hybrid vision/inertial tracking systems. *2nd InverVis 2005: Workshop on Integration of Vision and Inertial Systems*, Barcelona, Spain.

Lawrence, A. (1998). *Modern Inertial Technology: Navigation, Guidance, and Control*, 2nd edn. Mechanical Engineering Series. Springer, ISBN 0-387-98507-7.

Leone, G. (1998). The effect of gravity on human recognition of disoriented objects. *Brain Research Reviews*, **28**(1–2): 203–214.

Lobo, J. (2002). *Inertial Sensor Data Integration in Computer Vision Systems*. Master's thesis, University of Coimbra, Portugal.

Lobo, J. (2006). InerVis Toolbox for Matlab. http://www.deec.uc.pt/ jlobo/InerVis WebIndex/.

Lobo, J. and Dias, J. (2003). Vision and inertial sensor cooperation using gravity as a vertical reference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(12): 1597–1608.

Lobo, J. and Dias, J. (2004). Inertial sensed ego-motion for 3d vision. *Journal of Robotic Systems*, **21**(1): 3–12.

Lobo, J. and Dias, J. (2005). Relative pose calibration between visual and inertial sensors. *ICRA 2005 Workshop on Integration of Vision and Inertial Sensors (InerVis2005)*, Barcelona, Spain.

Lobo, J. and Dias, J. (2007). Relative pose calibration between visual and inertial sensors. *International Journal of Robotics Research, Special Issue from the 2nd Workshop on Integration of Vision and Inertial Sensors*, to appear.

Lobo, J., Queiroz, C., and Dias, J. (2003). World feature detection and mapping using stereovision and inertial sensors. *Robotics and Autonomous Systems*, **44**(1): 69–81.

Ma, Y., Soatto, S., Kousecká, J., and Sastry, S. S. (2004). *An Invitation to 3-D Vision*. Springer.

Molton, N. and Brady, M. (2000). Practical structure and motion from stereo when motion is unconstrained. *International Journal of Computer Vision*, **39**(1): 5–23.

Montemerlo, M. and Thrun, S. (2007). *The FastSLAM Algortihm for Simultaneous Localization and Mapping*. Springer.

Muratet, L., Doncieux, S., Briere, Y., and Meyer, J.-A. (2005). A contribution to vision-based autonomous helicopter flight in urban environments. *Robotics and Autonomous Systems*, **50**(4): 195–209.

Naimark, L. and Foxlin, E. (2002). Circular data matrix fiducial system and robust image processing for a wearable visioninertial self-tracker. *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR'02)*, IEEE Computer Society, P. 27.

Neumann, U., You, S., Hu, J., Jiang, B., and Lee, J. (2003). Augmented virtual environments (ave): dynamic fusion of imagery and 3D models. *Proceedings of IEEE Virtual Reality*, pp. 61–67.

Newman, P. (ed.) (2007). Special issue on SLAM in the field. *Journal of Field Robotics*, **24**(1,2).

Nistér, D. (2005). Preemptive ransac for live structure and motion estimation. *Machine Vision Applications*, **16**(5): 321–329.

Nistér, D., Naroditsky, O., and Berge, J. (2006). Visual odometry for ground vehicle applications. *Journal of Field Robotics*, **23**(1): 3–20.

Nygards, J., Skoglar, P., Ulvklo, M., and Hgstrm, T. (2004). Navigation aided image processing in uav surveillance: Preliminary results and design of an airborne experimental system. *Journal of Robotic Systems*, **21**(2): 63–72.

Panerai, F., Metta, G., and Sandini, G. (2000). Visuoinertial stabilization in space-variant binocular systems. *Robotics and Autonomous Systems*, **30**(1–2): 195–214.

Panerai, F., Metta, G., and Sandini, G. (2002). Learning visual stabilization reflexes in robots with moving eyes. *Neurocomputing*, **48**(1–4): 323–337.

Panerai, F. and Sandini, G. (1998). Oculo-motor stabilization reflexes: integration of inertial and visual information. *Neural Networks*, **11**(7–8): 1191–1204.

Previc, F. H., Liotti, M., Blakemore, C., Beer, J., and Fox, P. (2000). Functional imaging of brain areas involved in the processing of coherent and incoherent wide field-of-view visual motion. *Experimental Brain Research*, **131**(4): 393–405.

Qian, G., Chellappa, R., and Zheng, Q. (2001). Robust structure from motion estimation using inertial data. *Journal of the Optical Society of America A*, **18**(12): 2982–2997.

Qian, G., Chellappa, R., and Zheng, Q. (2002). Bayesian structure from motion using inertial information. *Proceedings of the International Conference on Image Processing*, Vol.3, pp. III–425–III–428.

Rehbinder, H. and Ghosh, B. (2003). Pose estimation using line-based dynamic vision and inertial sensors. *IEEE Transaxctions on Automatic Control*, **48**(2): 186–199.

Reymond, G., Droulez, J., and Kemeny, A. (2002). Visuovestibular perception of self-motion modeled as a dynamic optimization process. *Biological Cybernetics*, **87**(4): 301–314.

Ribo, M., Brandner, M., and Pinz, A. (2004). A flexible software architecture for hybrid tracking. *Journal of Robotic Systems*, **21**(2): 53–62.

Roetenberg, D., Luinge, H., Baten, C., and Veltink, P. (2005). Compensation of magnetic disturbances improves inertial and magnetic sensing of human body segment orientation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, [see also *IEEE Transactions on Rehabilitation Engineering*], **13**(3): 395–405.

Roetenberg, D., Luinge, H., and Veltink, P. (2003). Inertial and magnetic sensing of human movement near ferromagnetic materials. *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 268–269.

Roumeliotis, S. I., Johnson, A. E., and Montgomery, J. F. (2002). Augmenting inertial navigation with image-based motion estimation. *Proceedings of International Conference on Robotics and Automation*, IEEE, p. 4326.

Savage, P. G. (1984). *Strapdown System Algorithms, Advances in Strapdown Inertial Systems*, Chapter 3, pp. 3.1–3.30. Lecture Series 133. AGARD, Advisory Group for Aerospace Research and Development.

Soatto, S., Perona, P., Frezza, R., and Picci, G. (1993). Recursive motion and structure estimation with complete error characterization. *Proceedings of Conference on Computer Vision and Pattern Recognition*, New York, pp. 428–433.

Stratmann, I. and Solda, E. (2004). Omnidirectional vision and inertial clues for robot navigation. *Journal of Robotic Systems*, **21**(1): 33–39.

Strelow, D. (2004). *Motion Estimation from Image and Inertial Measurements*. PhD thesis, Carnegie-Mellon.

Strelow, D. and Singh, S. (2002). Optimal motion estimation from visual and inertial measurements. *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, IEEE Computer Society, p. 314.

Strelow, D. and Singh, S. (2003). Optimal motion estimation from visual and inertial measurements. *Proceedings of the Workshop on Integration of Vision and Inertial Sensors (INERVIS 2003)*.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press.

Tomasi, C. and Shi, J. (1994). Good features to track. *CVPR94*, pp. 593–600.

Torr, P. A structure and motion toolkit in matlab. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL COPIES/TORR1/index.html.

Triggs, B., McLauchlan, P., Hartley, R., and Fitzgibbon, A. (2000). Bundle adjustment – A modern synthesis. In *Vision Algorithms: Theory and Practice* (eds W. Triggs, A. Zisserman, and R. Szeliski, pp. 298–375, LNCS, Springer Verlag.

Viéville, T. (1997). *A Few Steps Towards 3D Active Vision*. Springer-Verlag, ISBN=3540631062.

Viéville, T., Clergue, E., and Facao, P. (1993a). Computation of ego-motion and structure from visual an inertial sensor using the vertical cue. *Proceedings of the Fourth International Conference on Computer Vision*, pp. 591–598.

Viéville, T., Clergue, E., and Facao, P. E. D. S. (1995). Computation of ego motion using the vertical cue. *Machine Vision Applications*, **8**(1): 41–52.

Viéville, T. and Faugeras, O. (1989). Computation of inertial information on a robot. In *Fifth International Symposium on Robotics Research* (eds H. Miura and S. Arimoto), pp. 57–65, MIT Press.

Viéville, T. and Faugeras, O. (1990). Cooperation of the inertial and visual systems. In *Traditional and NonTraditional Robotic Sensors* (ed. T. C. Henderson), Vol. F 63 of NATO ASI, pp. 339–350, SpringerVerlag, Berlin, Heidelberg.

Viéville, T., Romann, F., Hotz, B., Mathieu, H., Buffa, M., Robert, L., Facao, P., Faugeras, O., and Audren, J. (1993b). Autonomous navigation of a mobile robot using inertial and visual cues. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (eds M. Kikode, T. Sato, and K. Tatsuno), Vol. 1, pp. 360–367, Yokohama, Japan.

Vincze, M. and Hager, G. D. (1999). *Robust Vision for Vision-Based Control of Motion*. SPIE/IEEE Press.

Viollet, S. and Franceschini, N. (2005). A high speed gaze control system based on the vestibulo-ocular reflex. *Robotics and Autonomous Systems*, **50**(4): 147–161.

Wall, C., Merfelda, D., Raucha, S., and Blackf, F. (2003). Vestibular prostheses: The engineering and biomedical issues. *Journal of Vestibular Research*, **12**: 95–113.

XVision. The XVision 2 project. http://www.cs.jhu.edu/CIRL/XVision2/.

Yazdi, N., Ayazi, F., and Najafi, K. (1998). Micromachined inertial sensors. *Proceedings of the IEEE*, **86**(8): 1640–1659.

You, S. and Neumann, U. (2001). Fusion of vision and gyro tracking for robust augmented reality registration. *Proceedings of IEEE Virtual Reality*, IEEE Computer Society, p. 71.

You, S., Neumann, U., and Azuma, R. (1999). Orientation tracking for outdoor augmented reality registration. *IEEE Computer Graphics and Applications*, **19**(6): 36–42.

Zhang, Z., Deriche, R., Faugeras, O., and Luong, Q.-T. (1995). A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal*, **78**: 87–119.

Zimmerman, M. and Sulzer, W. (1991). High bandwidth orientation measurement and control ased on complementary filtering. *Proceedings of Symposium on Robotics and Control, SYROCO*, Vienna.