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 deadreckoning (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.

* Corresponding author
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.

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.
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.

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

\[
\ddot{x}(t) + 2\zeta \omega_n \dot{x}(t) + \omega_n^2 x(t) = -\ddot{y}(t)
\]

which converts acceleration of the sensor body, \( \ddot{y}(t) \), to displacement \( x(t) \) with a natural frequency \( \omega_n \) and a damping ratio \( \zeta \). 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
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

\[ \mathbf{F} = 2m\mathbf{\omega} \times \mathbf{v} \]

proportional to the velocity \( \mathbf{v} \) of the moving particle, the rotation rate of the rotating reference frame \( \mathbf{\omega} \) and the particle’s mass \( m \).
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 functions 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, $\mathbf{a}$, over time to estimate position. Assuming a set of accelerometers measuring acceleration along three orthogonal axis we have

$$ \mathbf{p} = \int \mathbf{p} \, dt = \int \int \mathbf{p} \, dt = \int \int \mathbf{a} \, dt $$ (2)
Fig. 7. Axes of the semicircular canals (taken with permission from Dickman 2006).

where \( \mathbf{p} \) is the position, \( \mathbf{v} \) the velocity, and \( \mathbf{a} \) 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

\[
x_{t+1} = \Phi x_t + N(0, Q)
\]

\[
y'_t = H x_t + N(0, R)
\]

where the state vector \( x \) comprises system pose and its derivatives in the navigation frame of reference. The observation \( y' = [a, \omega] \) 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 \( q \) can be written as

\[
q = q_0 + q_1 i + q_2 j + q_3 k = (q_0, \mathbf{q})
\]

where \( q_1, q_2 \) and \( q_3 \) are the components of the imaginary or vector part \( \mathbf{q} \) of the quaternion, \( i, j \) and \( k \) 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 \( i \) in complex numbers, are all square roots of \(-1\), and \( i^2 = j^2 = k^2 = -1 \). The magnitude of a quaternion is defined as

\[
\|\mathbf{q}\| = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2}.
\]

The complex conjugate \( \mathbf{q}^* \) of quaternion \( \mathbf{q} \) is given by

\[
\mathbf{q}^* = q_0 - q_1 i - q_2 j - q_3 k = (q_0, -\mathbf{q})
\]

and the inverse \( \mathbf{q}^{-1} \)

\[
\mathbf{q}^{-1} = \frac{1}{\mathbf{q}} = \frac{\mathbf{q}^*}{\mathbf{q} \mathbf{q}}.
\]

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

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

\[
\mathbf{q} = \cos \frac{\phi}{2} + \sin \frac{\phi}{2} \mathbf{u}.
\]  

(9)

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

\[
\mathbf{p}' = \mathbf{q}^{-1} \mathbf{p} \mathbf{q}.
\]  

(10)

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

\[
\mathbf{a}_{\text{nav}} = \hat{\mathbf{q}}(k) \mathbf{a}_b \hat{\mathbf{q}}(k)^*. 
\]  

(11)

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

\[
\mathbf{\omega} = \begin{bmatrix} \omega_x & \omega_y & \omega_z \end{bmatrix}^T
\]  

(12)

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

\[
\Delta \hat{\mathbf{q}} = \cos \left( \frac{\mathbf{\omega} \Delta t}{2} \right) - \sin \left( \frac{\mathbf{\omega} \Delta t}{2} \right) \frac{\mathbf{\omega}}{\|\mathbf{\omega}\|} 
\]  

(13)

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

\[
\hat{\mathbf{q}}(k+1) = \hat{\mathbf{q}}(k) \Delta \hat{\mathbf{q}} 
\]  

(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 \( \mathbf{x} = [\mathbf{p}, \dot{\mathbf{p}}, \ddot{\mathbf{p}}, \hat{\mathbf{q}}, \dot{\hat{\mathbf{q}}}] \) where \( \mathbf{p} \in \mathbb{R}^3 \) and \( \hat{\mathbf{q}} \in \text{SO}(3) \). 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 (Jazwinski 1970). The state vector may be extended to include sensor bias and scale factors.
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

\[
sp_i = \begin{bmatrix} su \\ sv \\ s \end{bmatrix} = CP
\]

\[
= \begin{bmatrix} f_u & 0 & 0 & 1 \\ 0 & f_v & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}
\]

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 diaphragm 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 \( \omega = [\omega_x, \omega_y, \omega_z] \) and translational velocity \( T = [T_x, T_y, T_z] \) with respect to the fixed frame and \( P \) is a point in the world. The velocity of the point \( P \), expressed relative to the camera frame, is given by

\[
\dot{P} = \omega \times P + T.
\]  

(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 \( P \) in terms of the image feature parameters \( u, v \) as

\[
\dot{X} = \frac{v_z}{\lambda} \omega_z - \frac{v_x}{\lambda} \omega_x + T_x
\]

\[
\dot{Y} = \frac{u_z}{\lambda} \omega_z - \omega_x + T_y
\]

\[
\dot{Z} = \frac{z}{\lambda} (v \omega_x - u \omega_y) + T_z
\]

(19)

(20)

(21)

where \( f = f_u = f_v \). Our visual feature is the image plane coordinate \([u, v]\) corresponding to \( P \), and using the quotient rule we obtain
\[ \dot{u} = \frac{f Z \dot{x} - x \dot{Z}}{Z^2} \]
\[ = \frac{f}{Z^2} \left\{ Z \left[ Z \omega_y - \frac{v z}{f} \omega_z + T_x \right] \right\} \]
\[ - \frac{u z}{f} \left\{ Z \left[ v \omega_x - u \omega_y + T_z \right] \right\} \]
\[ = \frac{f}{Z^2} T_x - \frac{u}{Z} T_z - \frac{u v}{f} \omega_x + f^2 + u z^2 - v \omega_z \]
\[ \frac{\dot{v}}{Z} = \frac{f}{Z} T_y - \frac{v}{Z} T_z + \frac{f^2 - v^2}{f} \omega_x + \frac{u v}{f} \omega_y + u \omega_z. \]  

Rewriting these two equations in matrix form we obtain
\[
\begin{bmatrix}
\dot{u} \\
\dot{v}
\end{bmatrix} =
\begin{bmatrix}
\frac{f}{Z} & 0 & -\frac{u}{Z} & \frac{uv}{f} & \frac{(-f^2 + u^2)}{f} & -v \\
0 & \frac{f}{Z} & -\frac{v}{Z} & \frac{(f^2 + u^2)}{f} & \frac{uv}{f} & u
\end{bmatrix}
\times
\begin{bmatrix}
T_x \\
T_y \\
T_z \\
\omega_x \\
\omega_y \\
\omega_z
\end{bmatrix}
\]  

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 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 detectors (Harris and Stephens 1988) 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 similarities of interest region 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 Vincez and Hager (1999), and implementations of trackers are generally available (Hager and Toyama 1998; Xvision2; Birchfield). SFM using lines is discussed in Rebhinder 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

\[
x_{t+1} = \Phi x_t + N(0, Q) \\
y^*_t = H(x_t) + N(0, R)
\]  

where the state vector \( x = [x^e | x^w] \) comprises the state of the camera \( x^e = [\mathbf{P}, \mathbf{\hat{q}}] \) 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 \[ \mathbf{y}_i = [(u_1, v_1), (u_2, v_2), \cdots (u_N, v_N)] \] 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 feature 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 superior 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.

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
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

\[
x_{i+1} = \Phi x_i + N(0, Q)
\]

(29)

\[
y^i_t = H^i(x_t) + N(0, R)
\]

(30)

\[
y^e_t = H^e(x_t) + N(0, R)
\]

(31)

where the state vector \( x = [x^v, x^w]^T \) comprises the inertial state of the camera \( x^v = [p, \dot{p}, \ddot{p}, q, \dot{q}] \) and the state of the world \( x^w = [P_1, P_2, \cdots, P_N] \), and the 3D coordinates of \( N \) scene points. The inertial observations \( y^i_t = [\dot{p}, \omega] \) are the outputs of the inertial sensors, while the visual observations, \( y^e_t \) 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.


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 Angekaki 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 real-time 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; Ribó 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


