EXPLOITING INERTIAL SENSING IN MOSAICING AND VISUAL NAVIGATION

Luiz G. B. Mirisola * **Jorge M. M. Dias** *

* *Institute of Systems and Robotics*
University of Coimbra, Portugal
{lgm,jorge}@isr.uc.pt

Abstract
In vision systems used in robotics, inertial and earth field magnetic sensors can provide valuable data about the observer ego-motion, as well as an absolute orientation reference. This article exploits the inertial orientation measurements to compensate the rotational degrees of freedom, in two different domains.
First, inertial data is used to project images on a leveled plane, relaxing the demands on interest point matching algorithms when performing image mosaicing. Second, in the rotation-compensated, pure translation case, full homographies are reduced to planar homologies, and the ratio of heights over the ground plane on two views are calculated more accurately. Both techniques are validated over outdoor image sequences including aerial images from an remotely piloted blimp.

Keywords: Vision, Inertial Measurement Units, Planes, Navigation.

1. INTRODUCTION

Vision systems in robotic applications can be rigidly coupled with Inertial Measurement Units (IMUs), which complement it with sensors providing direct measures of orientation relative to the world north-east-up frame, such as magnetometers and accelerometers (that measure gravity).

A novel calibration technique [Lobo and Dias, 2005] finds the rigid body rotation between the camera and IMU frames, and then the camera orientation in the world is obtained by rotating the IMU orientation measurement. The approximation of the rotational degrees of freedom should allow faster processing or the use of simpler movement models in computer vision tasks. For example, it can be explored to improve robustness on image segmentation and 3D structure recovery [Lobo et al., 2006].

The limits of computer vision or sensorial data fusion alone have already been largely explored, and it is known that some limits may be overcome by combining them.

In [Hygounenc et al., 2004], a stereovision-only aproach is used to build a 3D map of the environment from stereo images taken by a remotely controlled blimp, tracking the camera pose and landmarks on the ground. It was not their aim to integrate IMU measurements.

On-board inertial and GPS data, together with a dynamic model of the vehicle is used in [Brown and Sullivan, 2002] to project images taken from a high-flying airplane onto the ground plane. One-pixel accuracy is achieved with no need of image-based techniques.

Image mosaicing was performed in [Gracias, 2002], for an unmanned submarine navigating over flat sea-bottom, using only images as input. The registration converged only if the vehicle movement is restricted to be planar (no large change on roll and pitch).

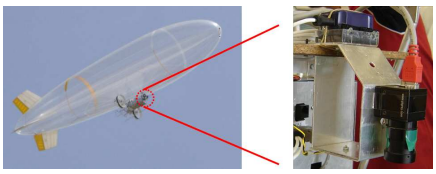

Figure 1. The vision-inertial system and an aerial vehicle that carries it.

Combined IMU and vision data were used to keep pose estimates in an underwater environment, navigating a robot submarine over a large area [Eustice, 2005]. Relative pose measurements from the images avoided divergence of the tracked vehicle pose, and an image mosaic is a byproduct.

In previous work [Mirisola et al., 2006] IMU sensed orientation aided the registration of stereo depth maps from a moving stereo camera. Each depth map was rotated to a leveled reference frame provided by the inertial sensed orientation. Then the remaining translation vector to register the 3D depth maps was found by interest point matching on the image sequence. A robust estimation process detects outliers from both interest point matching and stereo depth maps, and is very fast due to the simple translation vector model.

The aim of this article is to exploit the inertial orientation measurements in two other domains, separating rotational and translational components, and using simpler movement models that offer increased performance or accuracy.

In section 2 we discuss the registration of images over planar surfaces. As the camera orientation measurements allow us to rotate the stereo depth maps, images of the ground surface can also be registered into a common leveled plane, and be rotated to align with the north-east axes. In this way, the performance of interest point matching algorithms used in image mosaicing is shown to be improved.

Next, section 3 shows that in the rotation-compensated, pure translation case, planar homographies became homologies, a more restricted model that allows to calculate relative camera heights from pixel correspondences with more accuracy. Images from the UAV of figure 1 are used in the last experiment.

Finally, the conclusions are shown in section 4.

1.1 Definitions of reference frames

The camera provides intensity images I. The subscript i is the time index. Hence the following frames are defined, as shown in figure 2:

- **Camera Frame** $\{C\}$: This frame is used in the pinhole camera projection model. The origin is placed at the *camera center*, the axis z is the *depth*, and the axes x and y are parallel to the *image plane*. The camera is calibrated, its intrisic parameter matrix K is known, and f is its focal length.
- **Inertial Frame** $\{I\}$: The IMU outputs the rotation ${}^{W}R_{I}|_{i}$ from the $\{I\}$ to the $\{W\}$ frame.
- **World Frame** $\{W\}$: A NED (North East Down) frame.
- **Rotated Camera Frame** $\{R\}$: This virtual camera frame shares its origin with the $\{C\}$ frame, but its optical axes points in the direction of gravity, and the image axes are parallel to the north and east axes.

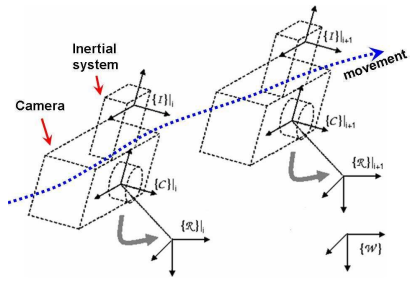

Figure 2. Definition of frames of reference.

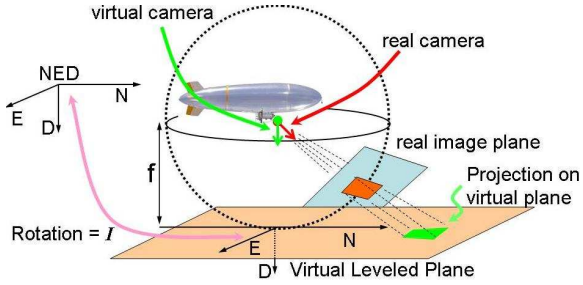

Figure 3. The virtual leveled plane concept.

The camera-inertial calibration outputs the constant rotation ${}^{I}R_{C}$ between the camera ($\{C\}$) and inertial ($\{I\}$) frames.

1.2 A virtual leveled plane

The knowledge of the camera orientation provided directly by the IMU measurements allows the image to be projected on entities defined on an absolute NED frame, such as a virtual horizontal plane (with normal parallel to gravity), at a distance f below the camera center, named as the *virtual leveled plane*, as shown in figure 3. Projection rays from 3D points to the camera center intersect this plane, projecting the 3D point into the plane. This projection corresponds to the image of a virtual camera at the $\{R\}$ frame, with optical axis coincident with the gravity vector. In the figure the moving observer is an UAV (out of scale).

1.3 Experimental Platforms

The moving observer hardware is shown in fig. 1. The camera is a Point Gray Flea [Point Gray Inc., 2006], and the inertial and magnetic sensor is a Xsens MT9-B [XSens Tech., 2006].

2. BUILDING IMAGE MOSAICS.

This section deals with the registration on the virtual leveled plane of an image sequence taken from a moving camera, rigidly coupled with an IMU. One arbitrary image is chosen as the reference image I_B, and the origin of its $\{R\}|_B$ frame is set as the origin of the $\{W\}$ frame.

2.1 Projecting on the virtual leveled plane.

For each image I_i, first the camera orientation in the $\{W\}$ frame is calculated as the rotation ${}^W R_C|_i = {}^W R_I|_i \cdot {}^I R_C$.

Then the image is transformed by the *infinite homography* [Ma et al., 2004], denoted by $H_\infty = K \cdot {}^R R_C|_i \cdot K^{-1}$. H_∞ is induced by the plane at infinity, i.e., it is the homography between two images taken from cameras at the same camera center, but rotated by the rotation matrix ${}^R R_C|_i$. Here it synthesizes a virtual view from a non-existent camera $\{R\}|_i$ with an image plane coincident with the virtual leveled plane - thus projecting the image on it. ${}^R R_C|_i = {}^R R_W \cdot {}^W R_C|_i$ is the rotation from the $\{C\}|_i$ to the $\{R\}|_i$ frame, where ${}^R R_W$ is a fixed rotation.

Any image can be picked as the reference one, as it is automatically projected to the desired mosaic plane orientation. In an image-only approach, the orientation of the mosaic plane must be retrieved from a specific image, or external inputs should be used.

2.2 Building mosaics with interest point matching

Once the images are projected into the virtual plane, interest point matching algorithms find pixel correspondences between pairs of such projected images, from which homographies are calculated to register these pairs. Small mosaics are built from successive overlapping frames in the sequence, registering as many frames as possible (two examples of small mosaics are shown in fig. 4, separately and then drawn together). Next the same algorithm is applied on the mosaics themselves, registering them into larger mosaics, and so on.

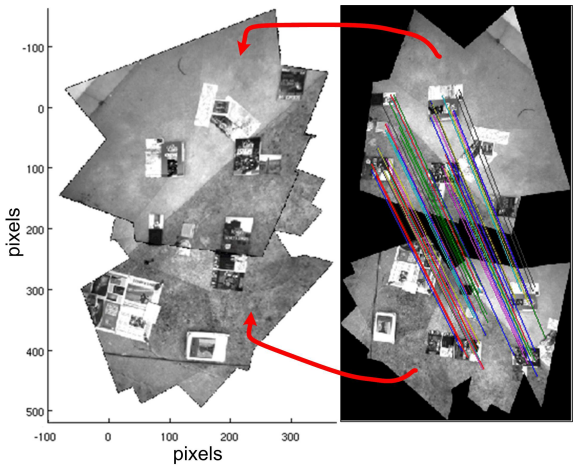

Figure 4. Right, two mosaics, with pixel correspondences; left, the mosaics registered.

The mosaic of fig. 5 was built from 61 images, taken from a tripod moved manually over a planar yard, with two different heights. From this sequence 20 mosaics were built, and then, a second run over them generated 4 larger mosaics. The final run generated the mosaic of fig. 5, with feathering to smooth image transitions.

This mosaic was obtained without deghosting and bundle adjustment, which usually must be applied to image-only mosaicing of this size [Szeliski, 2004], and still may be exploited to register larger datasets. Also, some recent results in mosaicing suppose a rotation-only model (e.g. [Brown and Lowe, 2003, Szeliski, 2004]), where the camera center is the same for all images. But here the camera is freely translating and rotating.

For comparison, the same interest point matching algorithms were applied to successive frames in the original image sequence as well as in the sequence projected into the virtual plane. The reprojection error (root mean square) on pixel correspondences was 20% less on the projected images.

Also, after tuning interest point detector parameters, a better ratio of number of matchings versus total number of interest points detected was obtained with the projected images, hence the matching of interest point descriptors was 50% faster, while still yielding a 2% larger number of correspondences.

3. CAMERA HEIGHTS FROM HOMOGRAPHIES AND HOMOLOGIES

Consider a 3D plane imaged in two views, and a set of pixel correspondences belonging to that plane, in the form of pairs of pixel coordinates $(\mathbf{x}, \mathbf{x}')$, representing the projection of the same 3D point on each view. The transformation relating these two sets of coordinates is a homography,

Figure 5. A mosaic from 61 registered images.

said to be *induced* by the plane. Given the two camera projection matrices $P = [I|\mathbf{0}]$ and $P' = [R|\mathbf{t}]$, the homography can be recovered from pixel correspondences [Ma et al., 2004], and it is related to the 3D plane normal $\mathbf{n}$, the distance from the camera center to the plane d, and the relative camera poses defined by a rotation matrix R and a translation vector $\mathbf{t}$, by:

$$\lambda H = \lambda \left(R - \mathbf{t}\mathbf{n}^T/d\right) \tag{1}$$

The module of the scale λ is the second largest singular value of λH , and the correct signal of H can be recovered by imposing a positive depth constraint. In the translation-only case, plane induced homographies become a special form called *planar homology*.

A planar homology G [van Gool et al., 1998] is a planar perspective transformation that has a line of fixed points (the *axis*), and another fixed point, the *vertex*. The axis is the image of the plane vanishing line (the intersection of the 3D plane and the plane at infinity), and the vertex is the epipole, or Focus of Expansion (FOE).

The cross ratios defined by the vertex, a pair of corresponding points, and the intersection of the line joining this pair with the axis, have the same value μ for all points. The matrix G is defined from the axis $\mathbf{a}$, vertex $\mathbf{v}$, and μ, by:

$$G = I + (\mu - 1)\frac{\mathbf{v}\mathbf{a}^T}{\mathbf{v}^T\mathbf{a}} \tag{2}$$

3.1 3D plane parallel to image plane

If the 3D plane is parallel to the image planes, the axis is the infinite line $\mathbf{a} = (0, 0, 1)^T$, and equation 2 becomes:

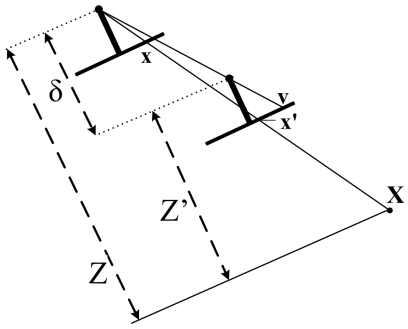

Figure 6. Two cameras under pure translation.

$$G = \begin{bmatrix} 1 & 0 & (\mu - 1) \cdot v_x \\ 0 & 1 & (\mu - 1) \cdot v_y \\ 0 & 0 & \mu \end{bmatrix} \tag{3}$$

where v_x, v_y are the *unhomogeneous* image coordinates of the vertex $\mathbf{v} = (v_x, v_y, 1)$. μ depends only of the relative depths of the 3D plane in both views. To analyze this relation, we recall that the relative scene depth of two points equals the reciprocal ratio of the image distances to the vanishing point of their connecting line [Arnspang et al., 1999].

This fact is true for two images of the same 3D point $\mathbf{X}$ under pure translation. Defining Z and Z' as the depth of $\mathbf{X}$ in first and second views, and $\mathbf{x}$ and $\mathbf{x}'$ as its respective projected image coordinates, as in fig. 6, we have:

$$\frac{Z'}{Z} = \frac{dist(\mathbf{x}, \mathbf{v})}{dist(\mathbf{x}', \mathbf{v})} \tag{4}$$

where *dist* is euclidean distance on the image. If a 3D plane is parallel to the image planes, all points on it have the same depth, and are transferred between the two views by the same homology.

The homology calculation involves many pairs of corresponding pixels, and thus is potentially more stable than an image measure involving just one pair. To relate the relative depth of the plane with the cross-ratio μ we recall that, given the homography matrix induced by a 3D plane in two views, the relative distance between the camera centers and the plane is equal to the determinant of the homography [Malis et al., 1999].

This is valid for full homographies, thus also for homologies. As, from equation 3, $det(G) = \mu$, and as the distance between the camera center and the plane is the depth of the plane, we have:

$$\frac{Z'}{Z} = \frac{dist(\mathbf{x}, \mathbf{v})}{dist(\mathbf{x}', \mathbf{v})} = \mu \tag{5}$$

3.2 Results: Relative Height for horizontal planes.

Again, rotation is compensated by projecting the images into the virtual leveled plane. In such

	rms error	std of error
full homography	0.055	0.036
homology	0.029	0.013

Table 1. Results for relative depth of 3D plane parallel to virtual image planes.

way, pure translation is simulated, and supposing that the camera views a flat horizontal plane, the camera height is equal to the plane depth. This section describes the process to calculate the ratio of the heights in two views.

A FOE estimate is obtained from pixel correspondences with outlier removal [Chen et al., 2003]. Then, from the pixel correspondences and the estimated FOE, μ is estimated by averaging the ratio of equation 5 for all corresponding pixel pairs.

Given the estimates for $\mathbf{v}$ and μ, an optimization routine minimizes the projection error of the correspondences when projected by the homology $G(\mathbf{v}, \mu, \mathbf{a} = [0, 0, 1]^T)$, finding improved estimates for $\mathbf{v}$ and μ. The relative depth is $det(G) = \mu$.

In the following experiment, the IMU-camera system of fig. 2 was mounted on a tripod, taking 50 images of the ground from different viewpoints. On this controlled environment the homology and homography models can be compared with hand-measured ground truth (this is not possible for the airship dataset used at the end of this section).

Figure 7 shows the height for all images, relative to the first image height ($104.5cm$). Two arrows connect two highlighted points to their respective images. The tripod was set to 3 different heights, thus the 3 horizontal lines are the ground truth. The stars are μ as described above. The crosses are the relative depths taken as the determinant of a full homography, estimated with *RANSAC*, optimized to minimize the projection error on pixel correspondences, and scaled as in eq. 1. The relative depths obtained from the full homography and from the homology model are compared, and the results, summarized in table 1, show that the latter offers improved accuracy.

Figure 8 shows a process diagram. There is no need to project all the image on the virtual plane, but only the coordinates of the pixel correspondences. Sensor data could provide directly an initial FOE estimate. The initial μ estimate is trivial, and the final optimization takes approximately as much time as the optimization for the homography. Therefore potentially this process can be fast enough for robotic applications.

In the following plane segmentation experiment, two images were taken from a staircase scene containing various horizontal planes, and image cross ratios were used to order the planes by their height. First, the image pair was projected into the leveled plane, pixel correspondences were

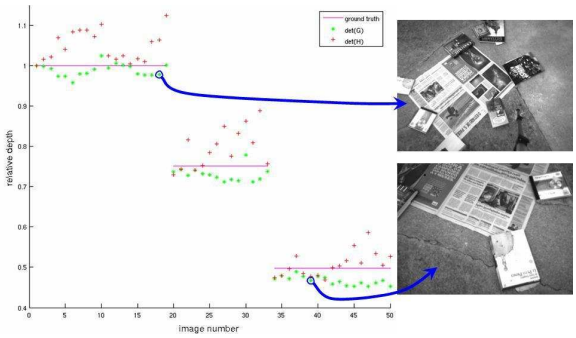

Figure 7. Relative heights to the ground from the tripod experiment, with two example images.

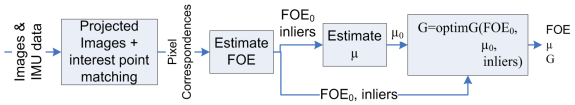

Figure 8. Finding the homology between two views.

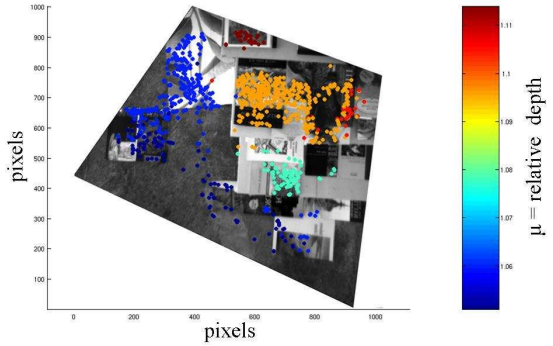

Figure 9. The relative depths from two views used to order planes by their height.

found, and from them, the FOE was calculated. Then, image cross ratios with the FOE were calculated for all corresponding pixel pairs, and groups of these pairs with close cross ratio values were found by picking the peaks of their histogram. Figure 9 shows each group with a different color, and the scale relating colors to relative depths is shown on the right. These points are very fast to obtain, and they can be seeds for plane segmentation algorithms.

The last result was obtained from images taken by the remote controlled blimp of fig. 1 carrying the IMU-camera calibrated system and GPS, flying over a planar area. The GPS measured height is shown in figure 10 compared against visual odometry based on the μ value of homologies calculated for the image sequence by the process described here. The height of the first image is manually set as $h_1 = 4m$, and the height of the ith image is $h_i = \left(\prod_{j=1}^{i-1} {}^{j+1}\mu_j\right) \cdot h_1$, where ${}^{j+1}\mu_j$ is the cross ratio of the homology that transforms the jth image into the image $j+1$. For the few image pairs where the homology could not be calculated, the last valid μ value was assumed

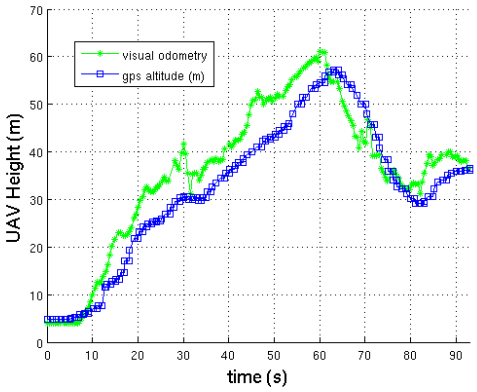

Figure 10. Visual odometry based on homology compared with GPS altitude measurements.

to be the current one. No other attempt was made to filter the data to avoid the drift from successive multiplication of relative heights. The scale depends on the manually set height h_1. As GPS altitude is not very accurate, the comparison only shows the existence of correlation.

The IMU orientation output was directly used, and its standard deviation for an static IMU (as on the tripod experiment) is 3° (it is larger for the moving UAV). Errors on the orientation increase the reprojection error of the homology, turning correctly matched pixel pairs into outliers, with a more significative effect as the distance from the optical center of the $\{R\}$ camera increases. But on these experiments there were still enough inliers for a reliable calculation.

4. CONCLUSION

Inertial orientation measurements and computer vision were combined in two different domains. The IMU data approximated the rotational degrees of freedom, and images were projected on an earth-grounded virtual plane. While image-only mosaicing is commonly done, this IMU-based projection improves the performance of essential parts of the mosaicing process.

The virtual plane projection also aids to determine relative heights, by simulating pure translation, and enabling the use of the homology model, that has been shown to be more accurate than full homographies. Encouraging results were shown, both in controlled laboratory environments where ground truth can be measured, and on aerial images taken from an UAV using directly the orientation output of an off-the-shelf IMU.

In further developments these ideas could be applied to 3D planes in general position, and the FOE could be directly measured. The height measurements could be used for landing aerial vehicles, or as an additional altitude sensor. The inertial-camera calibration is a key technique to make these ideas useful for robotic mapping and navigation, including aerial robotics.

REFERENCES

J. Arnspang, K. Henriksen, and F. Bergholm. Relating scene depth to image ratios. In *8th Int. Conf. on Computer Analysis of Images and Patterns (CAIP'99)*, pages 516–525, Ljubljana, Slovenia, Sep 1999.

A. Brown and D. Sullivan. Precision kinematic alignment using a low-cost GPS/INS system. In *ION GPS*, Portland, OR, USA, Sep. 2002.

M. Brown and D. G. Lowe. Recognising panoramas. In *10th Int. Conf. on Computer Vision (ICCV)*, Nice, France, October 2003.

Z. Chen, N. Pears, J. McDermid, and T. Heseltine. Epipole estimation under pure camera translation. In C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen, editors, *DICTA*, pages 849–858. CSIRO Publishing, 2003. ISBN 0-643-09041-X.

R. Eustice. *Large-Area Visually Augmented Navigation for AUV*. PhD thesis, Massachusetts Institute of Technology, June 2005.

N. Gracias. *Mosaic-based Visual Navigation for AUV*. PhD thesis, Instituto Superior Técnico, Lisbon, Portugal, December 2002.

E. Hygounenc, I-K. Jung, P. Soueres, and S. Lacroix. The Autonomous Blimp Project at LAAS/CNRS. *Int. J. of Robotics Research*, 23 (4/5):473–512, April/May 2004.

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

J. Lobo, J. F. Ferreira, and J. Dias. Bioinspired visuo-vestibular artificial perception system for independent motion segmentation. In *ICVW06 (2nd Int. Cognitive Vision Workshop)*, Graz, Austria, May 2006.

Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3D Vision*. Springer, 2004.

E. Malis, F. Chaumette, and S. Boudet. 2-1/2-D Visual Servoing . *IEEE Trans. on Robotics and Automation*, 15(2):238–250, April 1999.

L. Mirisola, J. Lobo, and J. Dias. Stereo vision 3D map registration for airships using vision-inertial sensing. In *12th IASTED Int. Conf. on Robotics and Applications (RA2006)*, Honolulu, HI, USA, August 2006.

Point Gray Inc., 2006. www.ptgrey.com.

R. Szeliski. Image alignment and stitching: A tutorial. Technical Report MSR-TR-2004-92, Microsoft Research, December 2004.

L. van Gool, M. Proesmans, and A. Zisserman. Planar homologies for grouping and recognition. *Image and Vision Computing*, 16(21-26), January 1998.

XSens Tech., 2006. www.xsens.com.