Tracking from a Moving Camera with Attitude Estimates.

Luiz G. B. Mirisola* and Jorge Dias

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

Abstract. This article presents a technique for object and people tracking using images acquired by a moving camera, which takes an image sequence of a planar leveled ground area where the target moves. Orientation measurements from an AHRS compensate the rotational degrees of freedom of the camera motion. Our previous work showed that with the rotation compensated the camera trajectory can be recovered more accurately with pure translation models. In this paper these gains are further exploited to improve the tracking accuracy. The images are registered in a common 2D frame of reference, by first reprojecting the images into a virtual horizontal plane, and then registering the images with 2D translation and scaling. Then, a target moving on the ground, seen in the reprojected image sequence, is tracked in a 2D frame of reference. Results from the method assessment are presented by using real dynamic scenes imaged by an airship navigating at low altitude, with comparison to GPS, and by using scenes from a urban people surveillance context.

1 Introduction

In our previous work [1,2], orientation measurements from an Attitude Heading Reference System (AHRS) compensated the rotational degrees of freedom of the motion of the remotely controlled airship of Fig. 1. Firstly, the images were reprojected in a geo-referenced virtual horizontal plane. Pure translation models were then used to recover the camera trajectory from images of a horizontal planar area, and they were found to be especially suitable for the estimation of the height component of the trajectory.

In this paper, the pure translation model with best performance is used to recover the camera trajectory while it images a target independently moving in the ground plane. The target trajectory is then recovered and tracked using only the observations made from a moving camera, including the airship on-board camera, as it is shown in Fig. 2(b), and results in a urban people surveillance context with known ground truth.

GPS also can be utilized to recover the airship trajectory, but GPS position fixes are less accurate in the altitude than in the latitude and longitude axes, and this uncertainty is very significant for the very low altitude dataset used in this paper.

Uncertainty in the camera orientation estimate is the most important source of error in tracking of ground objects imaged by an airborne camera [3], and its projection in the 2D ground plane is usually anisotropic even if the original distribution is isotropic. The Unscented Transform [4], which has been used to localize static targets on the ground [5], is thus used to project the uncertainty on the camera orientation estimate to the 2D ground plane, taking into account its anisotropic projection.

Kalman Filters are applied to the recovered trajectories of both camera and target. The latter trajectory is represented, tracked, and filtered in 2D coordinates. In this way the geometry of

* supported by the Portuguese Foundation for Science and Technology, grant BD/19209/2004

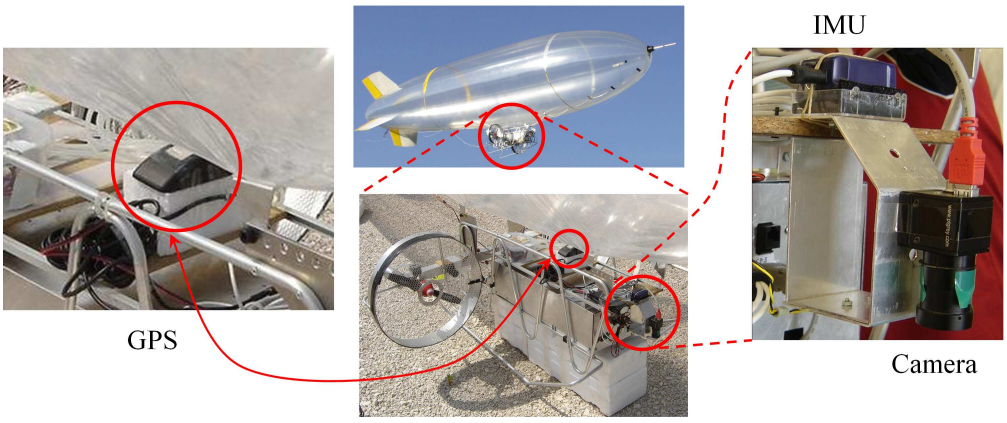


Fig. 1. An unmanned airship and detailed images of the vision-AHRS system and the GPS receiver mounted onto the gondola.

the camera and target motion is considered and the filters involved may utilize covariances and constants set accordingly to the camera and target motion in actual metric units and coordinate systems. This should allow for more accurate tracking than when only pixel coordinates in the images are utilized.

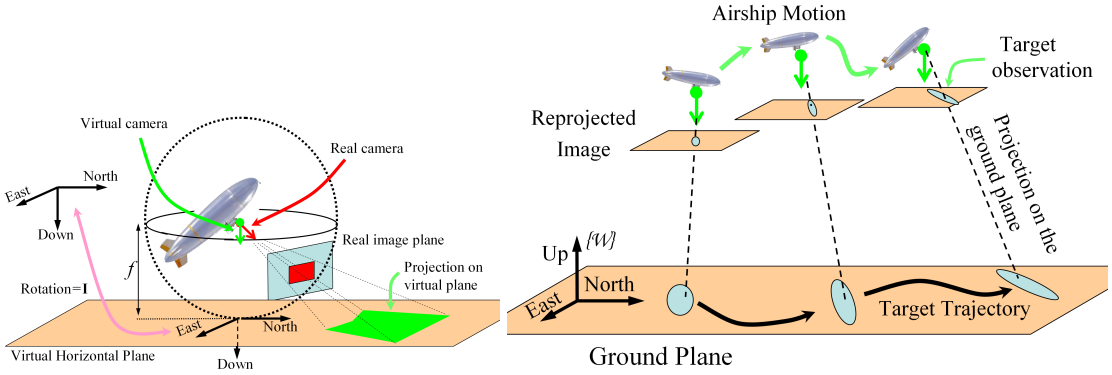
1.1 Experimental Platforms

The hardware used is shown in fig. 1. The AHRS used are Xsens MTi [6] for the airship experiment and a Xsens MTB-9 for the people tracking experiment. Both AHRS models use a combination of 3-axes accelerometers, gyroscopes and magnetic sensors to output estimates of their own orientation in geo-referenced coordinates. They output a rotation matrix ${}^{\mathcal{W}}\mathbf{R}_{AHRS}|_i$ which register the AHRS sensor frame with the north-east-up axes. The camera is a Point Gray Flea [7], with 1024×768 pixel resolution, capturing images at 5 fps. The camera is calibrated and the images corrected for lens distortion [8], its intrinsic parameter matrix $\mathbf{K}$ is known, and f is its focal length. To establish pixel correspondences in the images the SURF interest point library is used [9].

1.2 Definitions of Reference Frames

The camera provide intensity images $I(x, y)|_i$ where x and y are pixel coordinates and i is a time index. Besides the projective camera frame associated with the real camera (CAM) and the coordinate system defined by the measurement axes of the AHRS, the following other reference frames are defined:

- **World Frame** $\{\mathcal{W}\}$: A LLA (Latitude Longitude Altitude) frame, where the plane $z = 0$ is the ground plane. It is origin is an arbitrary point.
- **Virtual Downwards Camera** $\{\mathcal{D}\}|_i$: This is a projective camera frame, which has its origin in the center of projection of the real camera, but its optical axis points down, in the direction of gravity, and its other axes (i.e., the image plane) are aligned with the north and east directions.



(a) The virtual horizontal plane concept. (b) Target observations projected in the ground plane.

Fig. 2. Tracking an independently moving target with observations from a moving camera.

1.3 Camera-AHRS Calibration and a Virtual Horizontal Plane

The camera and AHRS are fixed rigidly together and the rotation between both sensor frames ${}^{AHRS}\mathbf{R}_{CAM}$ is found by the Camera Inertial Calibration Toolkit [10]. The translation between both sensors frames is considered negligible. The AHRS estimates of its own orientation are then used to estimate the camera orientation as ${}^{\mathcal{W}}\mathbf{R}_{CAM}|_i = {}^{\mathcal{W}}\mathbf{R}_{AHRS}|_i \cdot {}^{AHRS}\mathbf{R}_{CAM}$.

The knowledge of the camera orientation allows the images to be projected on entities defined on an absolute NED (North East Down) frame, such as a *virtual horizontal plane* (with normal parallel to gravity), at a distance f below the camera center, as shown in Fig. 2(a). 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 such as defined in Sect. 1.2. It is performed by the infinite homography [11], which depends on the calculation of the rotation between the real and virtual camera frames: ${}^{\mathcal{D}}\mathbf{R}_{CAM}|_i = {}^{\mathcal{D}}\mathbf{R}_{\mathcal{W}} \cdot {}^{\mathcal{W}}\mathbf{R}_{CAM}|_i$, where the rotation between the $\{\mathcal{D}\}|_i$ and $\{\mathcal{W}\}$ frames is defined as:

$$ {}^{\mathcal{D}}\mathbf{R}_{\mathcal{W}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{1} $$

1.4 Recovering the Camera Trajectory with a Pure Translation Model

Suppose a sequence of aerial images of a horizontal ground patch, and that these images are reprojected on the virtual horizontal plane as presented in Sect. 1.3. Corresponding pixels are detected between each image and the next one in the temporal sequence. The virtual cameras have horizontal image planes parallel to the ground plane. Then, each corresponding pixel is projected into the ground plane, generating a 3D point, as shown in Fig. 3(a). Two sets of 3D points are generated for two successive views, and these sets are directly registered in scene coordinates. Indeed, as all points belong to the same ground plane, the registration is solved in 2D coordinates. Figure 3(b) shows a diagram of this process.

Each corresponding pixel pair $(\mathbf{x}, \mathbf{x}')$ is projected by equation (2) yielding a pair of 3D points $(\boldsymbol{X}, \boldsymbol{X}')$, defined in the $\{\mathcal{D}\}|_i$ frame:

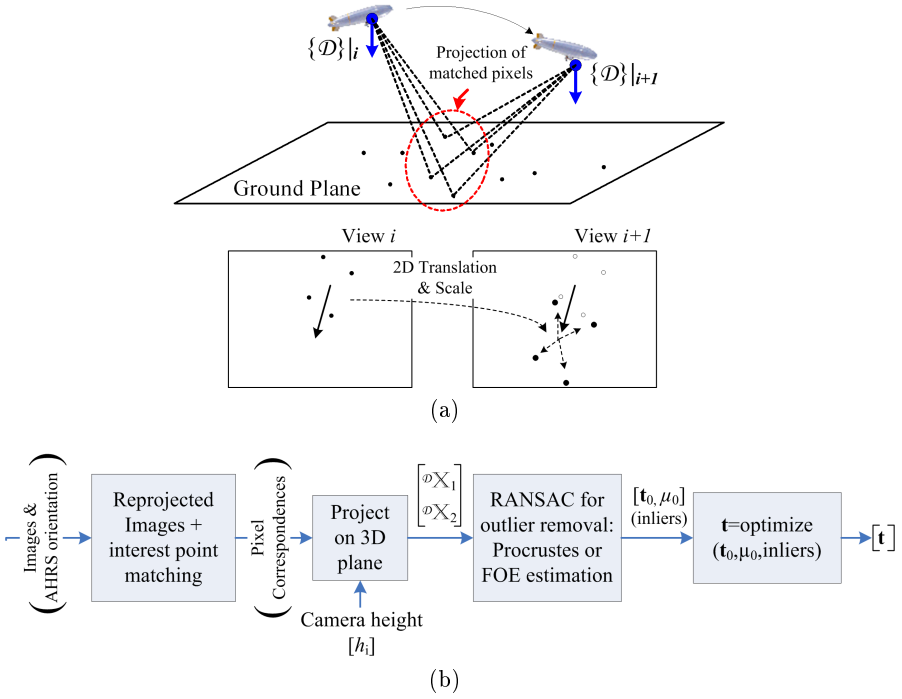


Fig. 3. Finding the translation between successive camera poses by 3D scene registration.

$$\boldsymbol{X} = \begin{bmatrix} \frac{(x_x - n_x)\cdot h_i}{f} \\ \frac{(x_y - n_y)\cdot h_i}{f} \\ h_i \end{bmatrix}, \quad \boldsymbol{X}'(\mathbf{t}) = \begin{bmatrix} \frac{(x'_x - n_x)\cdot (h_i - t_z/t_w)}{f} + \frac{t_x}{t_w} \\ \frac{(x'_y - n_y)\cdot (h_i - t_z/t_w)}{f} + \frac{t_y}{t_w} \\ h_i - t_z \end{bmatrix} \tag{2}$$

where $\mathbf{x} = [x_x, x_y, 1]^T$, $\mathbf{x}' = \left[x'_x, x'_y, 1\right]^T$, again in inhomogeneous form, h is the camera height above the ground plane, $\mathbf{t}$ is defined as a four element homogenous vector $\mathbf{t} = [t_x, t_y, t_z, t_w]^T$. The $\mathbf{t}$ value which turns $\boldsymbol{X}'(\mathbf{t}) = \boldsymbol{X}$ is the translation which registers the $\{\mathcal{D}\}|_i$ and $\{\mathcal{D}\}|_{i+1}$ frames, and which must be determined. If there are n corresponding pixel pairs, this projection yields two sets of 3D points, $\mathbb{X} = \{\boldsymbol{X}_k | k = 1 \ldots n\}$ and $\mathbb{X}' = \left\{\boldsymbol{X}'_k | k = 1 \ldots n\right\}$

An initial, inhomogeneous, value for $\mathbf{t}_0$ is calculated by the *Procrustes* registration routine [12]. It finds the 2D translation and scale factor which register the two point sets taken as 2D points, yielding estimates the x and y components of $\mathbf{t}_0$ and of the scale factor μ_0. The inputs for the Procrustes routine are the configurations $\mathbb{X}$ and $\mathbb{X}'(\mathbf{0})$.

From μ_0 and the current estimate of the camera height an initial estimate the vertical component of $\mathbf{t}_0$ can be calculated, as $\mu_0 = (h_i - t_z)/h_i$. Outliers in the pixel correspondences are removed by embedding the Procrustes routine in a RANSAC procedure. Then $\mathbf{t}_0$ is used as an initial estimate for an optimization routine which minimizes the registration error between $\mathbb{X}$ and $\mathbb{X}'(\mathbf{t})$, estimating an updated and final value for $\mathbf{t}$.

This optimization variables are the four elements of $\mathbf{t}$, with equation (2) used to update $\mathbb{X}'(\mathbf{t})$. The function to minimize is:

$$\min_{(t_x,t_y,t_z,t_w)} \sum_{k=1\ldots n} dist\left(\boldsymbol{X}'_k\left(\mathbf{t}\right), \boldsymbol{X}_k\right) \tag{3}$$

where *dist* is Euclidean distance.

The same process could be performed with an inhomogeneous $\mathbf{t}$. But, as it is the case with homography estimation, the over-parameterization improves the accuracy of the final estimate.

For datasets where the error in the orientation estimate is less significant, the algebraic Procrustes procedure obtains good results alone, with no optimization at all. Indeed, if the assumptions of having both image and ground planes parallel and horizontal are really true, with outliers removed, and considering isotropic error in the corresponding pixel coordinates, then it can be proved that the Procrustes solution is the best solution in a least squares sense. But the optimization step should improve robustness and resilience to errors, outliers and deviations from the model, and still exploit the available orientation estimate to recover the relative pose more accurately than an image-only method. More details and other pure translation models are shown in [1,2].

1.5 Filtering the Camera Pose

The camera trajectory is recovered as a sequence of translation vectors $\mathbf{t}$, considered as velocity measurements which are filtered by a Kalman Filter with a Wiener process acceleration model [13]. The filter state contains the camera position, velocity and acceleration. The process error considers a maximum acceleration increment of $0.35\,{}^{m}/_{s^2}$, and the sequence of translation vectors is considered as a measurement of the airship velocity, adjusted by the sampling period of 0.2 s. The measurement error is considered as a zero mean Gaussian variable with standard deviation of $1\,{}^{m}/_{s}$ in the vertical and $4{}^{m}/_{s}$ in the horizontal axes.

2 Tracking a Moving Target

Once the camera pose is known, a moving target is selected on each reprojected image. Problems such as image segmentation or object detection are out of the scope of this paper. Nevertheless, to track its position on the plane, the target coordinates on the virtual image must be projected on the $\{\mathcal{W}\}$ frame, considering the error in the camera position and orientation. Figure 4 summarizes this process which is detailed in this section.

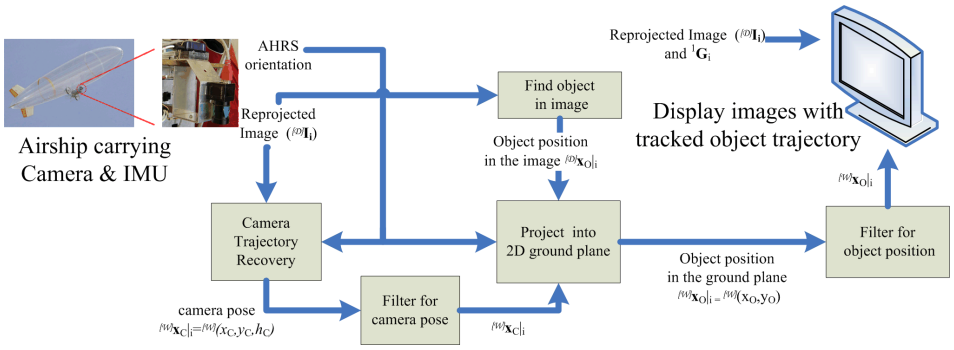


Fig. 4. A block diagram of the tracking process.

2.1 Target Pose Measurement: Projecting from Image to World Frame

The target coordinates in the image are projected into the ground plane by equation (2), and then these coordinates are transformed into the $\{\mathcal{W}\}$ frame by the appropriate rotation - equation (1) - and translation (the origin of the $\{\mathcal{D}\}$ frame is ${}^{\mathcal{W}}\mathbf{x}_C$ in the $\{\mathcal{W}\}$ frame).

The projection of the images in the virtual horizontal plane does not by itself improves the measurement of the target position on the ground, although it facilitates interest point matching [14,15]. The measurement of the position of an imaged target on the ground is very sensitive to errors in the camera orientation [3]. Therefore the uncertainty of the camera 6D pose is propagated with the Unscented Transform [5,4]. The actual errors in the camera position and orientation are unknown. The covariances found by the KF of Sect. 1.5 are used for the camera pose, and the camera orientation estimate is supposed to have zero mean Gaussian error with standard variation of 5°.

Therefore, given a sequence of camera poses with the respective images and an object detected on these images, this projection generates a sequence of 2D coordinates with anisotropic covariance ellipses for the target pose on the ground plane.

2.2 Filtering of Target Pose

The target pose is tracked in the 2D reference frame, and filtered by a Kalman Filter similar to the filter described in Sect. 1.5. The process error considers a maximum acceleration increment of $1\,{}^{m}/_{s^2}$, and the Unscented Transform supplies measurements of the target pose with covariance matrices which are considered as the measurement error. The target observations projected in the ground plane have high frequency noise, due to errors in the camera pose estimate and in the target detection in each image, thus the original target trajectory is filtered by a low pass filter with cut frequency of 2 Hz and attenuation of -10 dB before the input of the Kalman Filter.

3 Results

3.1 Tracking of a Moving Target from Airship Observations

Firstly, an object of known dimensions in the ground was observed, and the height of the camera estimated from its image dimensions, eliminating the scale ambiguity inherent to relative pose recovery from images alone. This was done a few seconds before the images shown. Then the airship trajectory was recovered by the model of Sect. 1.4. Only the Procrustes procedure was necessary as the optimization did not improve the results.

Fig. 5(a) shows the recovered airship trajectories using the method of Sect. 1.4 (red circles) and by the standard homography estimation and decomposition method (green crosses). The GPS trajectory is shown as blue squares and the target trajectory as blue stars.

The trajectories recovered by the method of Sect. 1.4 are shown again in Figure 5(b). The images projected in the ground plane by using equation (2) to find the coordinates of their corners in the ground plane and drawing the image in the canvas accordingly.

Figure 6 shows the a 2D view of the target trajectory over the corresponding images for the pure translation (a) and image-only (b) methods. The error in height estimation for the image-only method is apparent in figure 6(b) as an exaggeration in the size of the last images.

3.2 Tracking People with a Moving Surveillance Camera

The method described in Sect. 2 was applied to track a person moving on a planar yard, imaged by a highly placed camera which is moved by hand. The camera trajectory was recovered by the

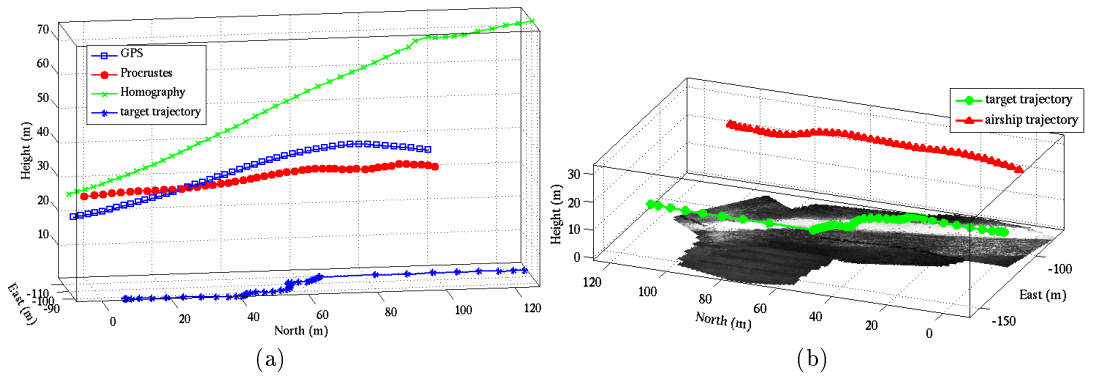


Fig. 5. A 3D view of the recovered trajectories: (a) Airship trajectories from GPS, pure translation and image-only method. Target trajectory derived from pure translation airship trajectory. (b) Trajectories recovered by the pure translation method, with registered images drawn on the ground plane.

method of Sect. 1.4 with no recourse to homography estimation. The square tiles in the floor provide ground truth, as the person was asked to walk only on the lines between squares. The trajectories of the camera and the target person are highlighted in Fig. 7(a), and Fig. 7(b) shows the recovered trajectories with the registered images in the top. The camera height above the ground was around 8.6 m, and each floor square has 1.2 m.

Figure 8(a) shows the recovered target trajectory to be compared with Fig. 8(b). In the latter case, the camera trajectory was recovered by the homography model. The red ellipses are 1 standard deviation ellipses for the covariance of the target position as estimated by the KF. In both figures, the large covariances in the bottom right appear when the target was out of the camera field of view, and therefore its estimated position covariance increased. When the target came back in the camera field of view the tracking resumed.

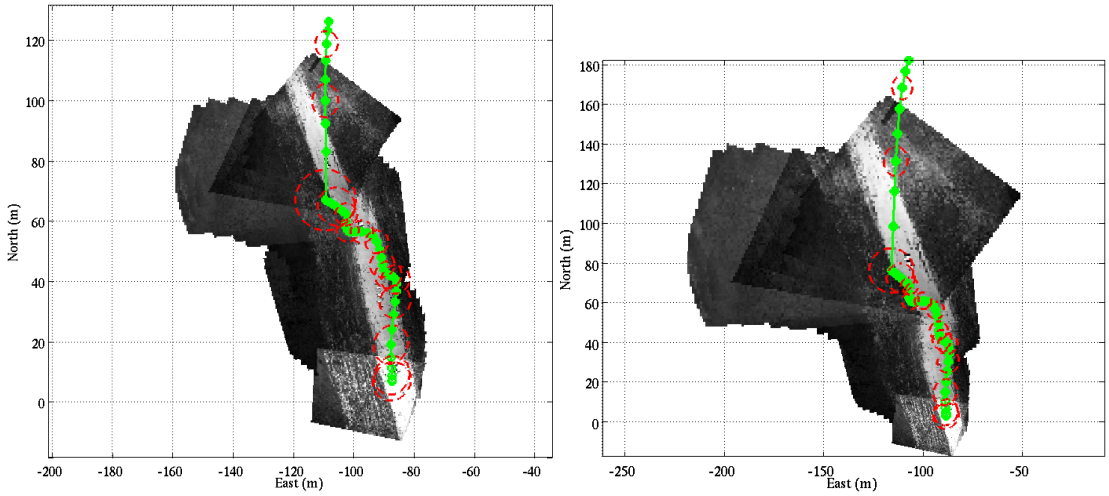
The solid yellow lines are the ground truth, marked over the floor images. Comparing the shape of the tracked trajectories is more significant than just the absolute difference to the ground truth, as the image registration itself has errors. The tracked trajectory after recovering the camera trajectory with the pure translation model appears more accurate than when the homography model is used.

4 Conclusion and Future Work

Our previous work on camera trajectory recovery with pure translation models was extended, with the same images being used to recover the moving camera trajectory and to track an independently moving target in the ground plane. The better accuracy of the camera trajectory recovery, or of its height component, resulted in better tracking accuracy. The filtering steps were performed in the actual metric coordinate frame instead of in pixel space.

With a low altitude UAV, GPS uncertainty is very significant, particularly as uncertainty in its altitude estimate is projected as uncertainty in the position of the tracked object, therefore recovering the trajectory from visual odometry can reduce the uncertainty of the camera pose, specially in the height component, and thus improve the tracking performance.

In the urban surveillance context these methods could be applied to perform surveillance with a camera carried by a mobile robot, extending the coverage area of a network of static cameras.

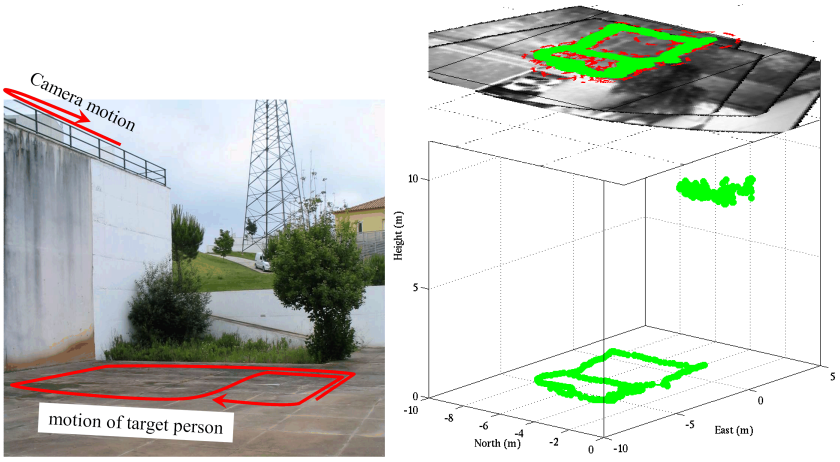


(a) Pure translation method. (b) Image only method.

Fig. 6. Tracking a car from the airship with the pure translation and the image only methods. The green circles are the tracked target trajectory with one standard deviation ellipses drawn in red.

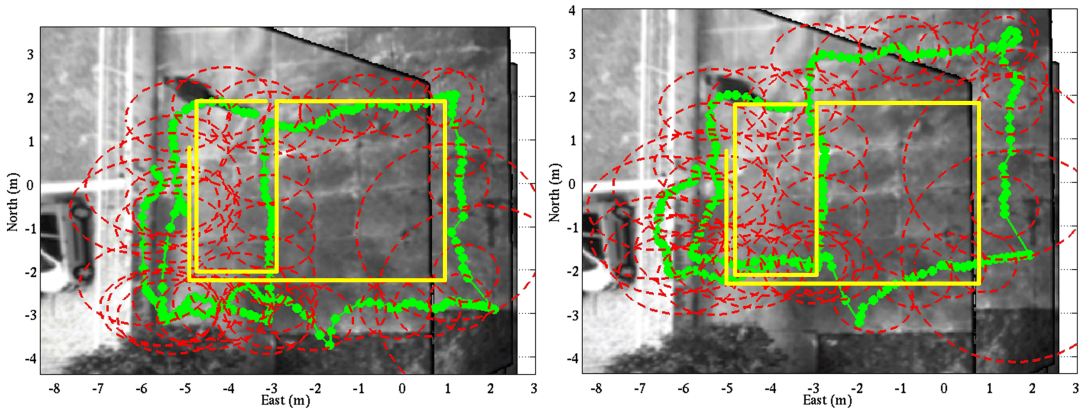
References

1. Mirisola, L.G.B., Dias, J.: Trajectory recovery and 3d mapping from rotation compensated imagery for an airship. In: Int. Conf. on Robots and Systems, San Diego, CA, USA (Nov. 2007)
2. Mirisola, L.G.B., Dias, J.: Exploting inertial sensing in vision based navigation with an airship. Journal of Field Robotics (2008) (submitted for publication).
3. Redding, J., McLain, T., Beard, R., Taylor, C.: Vision-based target localization from a fixed-wing miniature air vehicle. In: American Control Conference, Minneapolis, MN, USA (June 2006)
4. Julier, S.J., Uhlmann, J.K.: A new extension of the kalman filter to nonlinear systems. In: Int. Symp. Aerospace/Defense Sensing, Simul. and Controls, Orlando, FL, USA (1997)
5. Merino, L., Caballero, F., de Dios, J., Ollero, A.: Cooperative fire detection using unmanned aerial vehicles. In: IEEE Int. Conf. on Robotics and Automation, Barcelona, Spain (Apr. 2005) 1896–1901
6. XSens Tech.: (2007) www.xsens.com.
7. Point Grey Inc.: (2007) www.ptgrey.com.
8. Bouguet, J.: Camera Calibration Toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html (2006)
9. Bay, H., Tuytelaars, T., van Gool, L.: SURF: Speeded Up Robust Features. In: the Ninth European Conference on Computer Vision, Graz, Austria (May 2006)
10. Lobo, J., Dias, J.: Relative pose calibration between visual and inertial sensors. International Journal of Robotics Research **26**(6) (2007) 561–575
11. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, UK (2000)
12. Gower, J.C., Dijksterhuis, G.B.: Procrustes Problems. Oxford Statistical Science Series. Oxford University Press (2004)
13. Bar-Shalom, Y., Li, X.R., Kirubarajan, T.: Estimation with Applications to Tracking and Navigation. John Willey & Sons, Inc (2001)
14. Mirisola, L.G.B., Dias, J.: Exploiting inertial sensing in mosaicing and visual navigation. In: 6th IFAC Symp. on Intelligent Autonomous Vehicles, Toulouse, France (Sep 2007)
15. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., van Gool, L.: A comparison of affine region detectors. Int. J. of Computer Vision **65**(1-2) (Nov. 2005) 43–72 http://dx.doi.org/10.1007/s11263-005-3848-x.



(a) (b)

Fig. 7. A photo with highlighted trajectories of camera and target person (a). A 3D view of the recovered trajectories, using the method of Sect. 1.4 to recover the camera trajectory (b).



(a) Camera trajectory recovered by the method of Sect. 1.4

(b) Camera trajectory recovered by the homography model.

Fig. 8. A closer view of the target person tracked trajectory.