Chapter Number¶(18pt)

Tracking a Moving Target from a Moving Camera with Rotation-Compensated Imagery

Luiz G. B. Mirisola, Jorge Dias
Institute of Systems and Robotics - University of Coimbra
Portugal

1. Introduction

In our previous work [Mirisola and Dias, 2007b, Mirisola and Dias, 2008], 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 and the AHRS estimated orientation, including the camera and AHRS onboard the airship, as it is shown in Fig. 2(b), and results in a urban people surveillance context with known ground truth. To compare our pure translation method with an image-only method, the camera trajectory is also recovered by the usual homography estimation and decomposition method, and the target is also tracked from the corresponding camera poses.

GPS also can be utilized to recover the airship trajectory, but GPS position fixes are notoriously 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 [Redding et al., 2006], and its projection in the 2D ground plane is usually anisotropic even if the original distribution is isotropic. The Unscented Transform [Julier and Uhlmann, 1997], which has been used to localize static targets on the ground [Merino et al., 2005], 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 utilized to filter the recovered trajectories of both camera and the tracked target. In the airship scenario, the visual odometry and GPS position fixes can be fused together by the Kalman Filter to recover the airship trajectory. The target trajectory is represented, tracked, and filtered in 2D coordinates. In this way the full geometry of the camera and target motion is considered and the filters involved may utilize covariances and

constants set to the physical limits of 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.

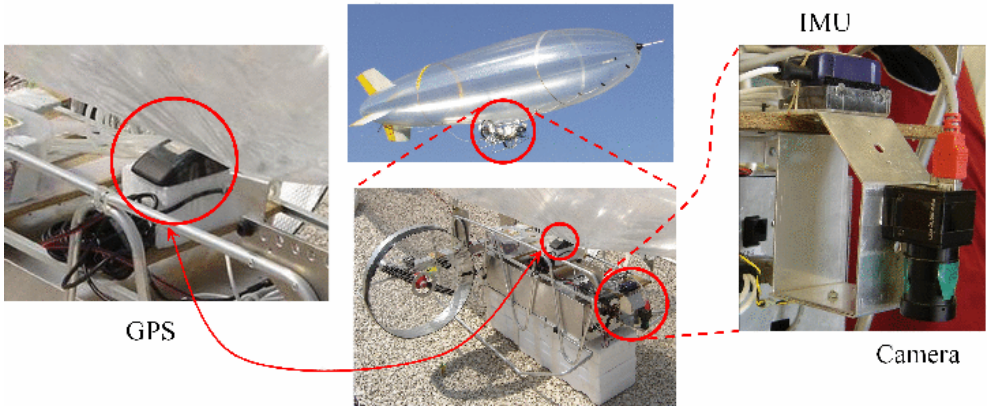


Fig. 1. An unmanned airship and detailed images of the vision-AHRS system and the GPS receiver mounted onto the gondola.

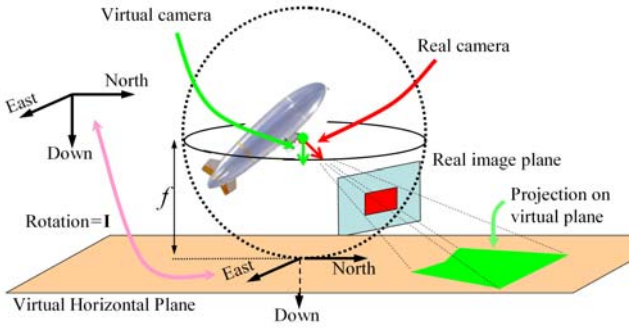
1.1 Experimental Platforms

The hardware used is shown in fig. 1. The AHRS used are Xsens MTi [XSens Tech., 2007] 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 ${}^{W}\mathbf{R}_{AHRS}|_i$ which registers the AHRS sensor frame with the north-east-up axes. The camera is a Point Gray Flea [Point Grey Inc., 2007], which captures images with resolution of 1024 × 768 pixels, at 5 fps. The camera is calibrated and its images are corrected for lens distortion [Bouguet, 2006], 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 [Bay et al., 2006].

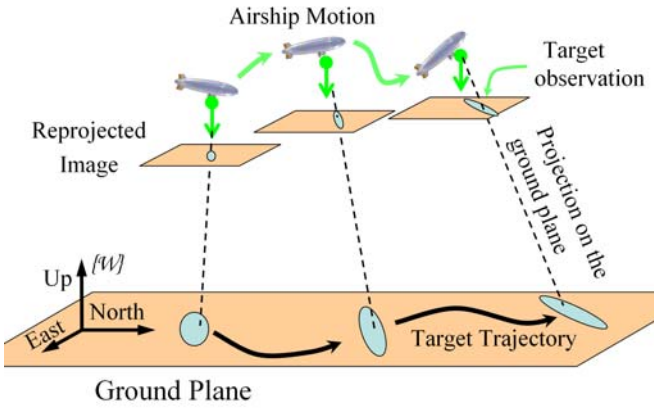
1.2 Definitions of Reference Frames

The camera provide intensity images $\mathbf{I}(x, y)|_i$ 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 centre 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

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 constant rotation between both sensor frames ${}^{AHRS}\mathbf{R}_{CAM}$. is found by the Camera-Inertial Calibration Toolkit [Lobo and Dias, 2007].

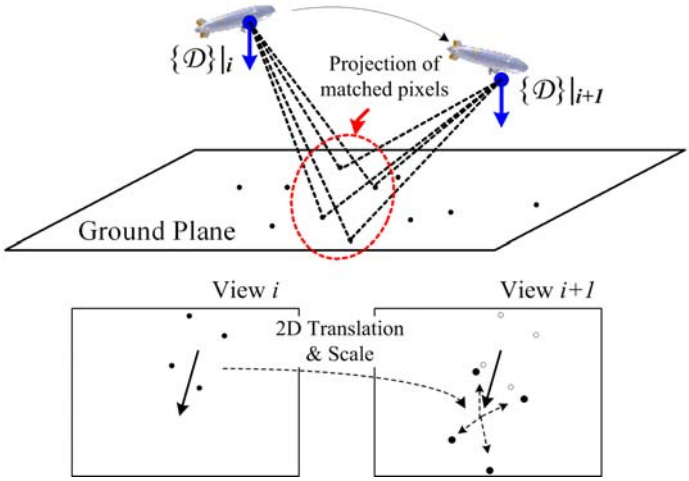
The translation between both sensors frames is negligible and considered as zero. The AHRS estimates of its own orientation are then used to estimate the camera orientation as ${}^{w}\mathbf{R}_{CAM}|_i = {}^{w}\mathbf{R}_{AHRS}|_i \cdot {}^{AHRS}\mathbf{R}_{CAM}$. The knowledge of the camera orientation 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 centre 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 [Hartley and Zisserman, 2000], 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, by definition, given by:

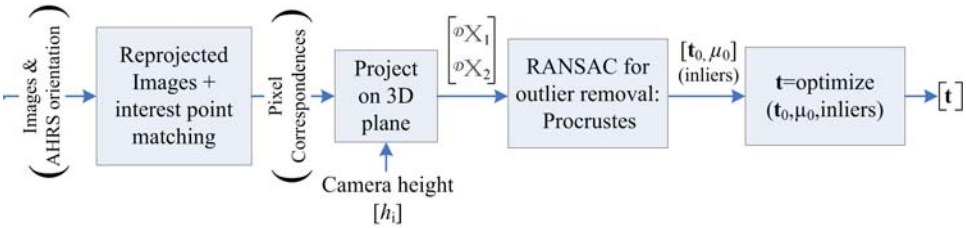
$$ {}^{\mathcal{D}}\mathbf{R}_{\mathcal{W}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{1} $$

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 section 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 figure 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.



(a)



(b)

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

Each corresponding pixel pair $(\mathbf{x}, \mathbf{x}_0)$ is projected by equation (2) yielding a pair of 3D points $(\mathbf{X},\mathbf{X}')$, defined in the $\{\mathcal{D}\}|_i$ frame:

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

where $\mathbf{x} = [x_x, x_y, 1]^T$, $\mathbf{x}' = [x'_x, x'_y, 1]^T$, 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 $\mathbf{X}'(\mathbf{t}) = \mathbf{X}$ 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, $\boldsymbol{\mathcal{X}} = \{\mathbf{X}_k \,|\, k = 1 \ldots n\}$ and $\boldsymbol{\mathcal{X}}' = \{\mathbf{X}'_k \,|\, k = 1 \ldots n\}$.
An initial, inhomogeneous, value for $\mathbf{t}_0$ is calculated by the Procrustes registration routine [Gower and Dijksterhuis, 2004]. 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 $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{X}}'(\mathbf{0})$.
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 $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{X}}'(\mathbf{t})$, estimating an updated and final value for $\mathbf{t}$.
The optimization variables are the four elements of $\mathbf{t}$, with equation (2) used to update $\boldsymbol{\mathcal{X}}'(\mathbf{t})$. The function to minimize is:

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

The same process could be performed with an inhomogeneous, three element $\mathbf{t}$. But, as it is the case with homography estimation, the over-parameterization improves the accuracy of the final estimate and sometimes even the speed of convergence. In this case the extra dimension allows the length of the translation to change without changing its direction.
For datasets where the actual camera orientation is almost constant or 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 [Mirisola and Dias, 2007b, Mirisola and Dias, 2008]

1.5. Filtering the Camera Pose

The camera trajectory is recovered as a sequence of translation vectors t, considered as velocity measurements which are filtered by a Kalman Filter with a Wiener process acceleration model [Bar-Shalom et al., 2001]. The filter state contains the camera position, velocity and acceleration. The filter should reduce the influence of spurious measurements and generate a smoother trajectory. The process error considers a maximum acceleration increment of 0.35 m/s², 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 4 m/s in the horizontal axes and 1 m/s in the vertical axis. The camera pose ${}^{w}\mathbf{X}_C(i)$ is taken from the filter state after the filtering.

2. Tracking of Moving Targets

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 reference $\{\mathcal{W}\}$ frame, considering the error in the camera position and orientation. Figure 4 summarizes this process which is detailed in this section.

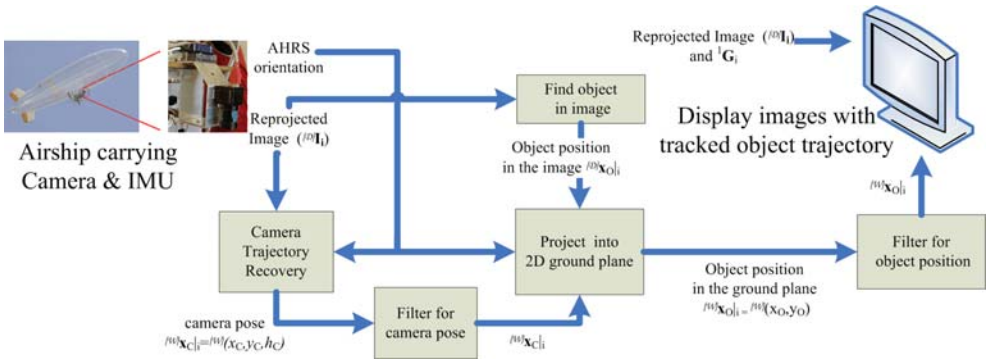


Fig. 4. A block diagram of the tracking process

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 ${}^{w}\mathbf{x}_C$ in the $\{\mathcal{W}\}$ frame).

The actual generation of reprojected images in the virtual horizontal plane does not by itself improve the measurement of the target position on the ground. Interest point matching could be performed with the original images, and only the coordinates of the matched interest points need to be actually reprojected on the virtual horizontal plane in order to apply the pure translation method. Nevertheless, interest point matching is faster or more robust if the rotation-compensated images are used [Mirisola and Dias, 2007a, Mikolajczyk

et al., 2005], and the reprojected images can be drawn on the ground plane forming a map as in figures 5 and 6.
The measurement of the position of a target imaged on the ground is known to be very sensitive to errors in the camera orientation [Redding et al., 2006]. Therefore the uncertainty of the camera 6D pose is propagated with the Unscented Transform, which has already been used to estimate the position of static targets observed from a low altitude UAV [Merino et al., 2005, Redding et al., 2006]. The actual errors in the camera position and orientation are unknown. The covariance found by the KF of section 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^o.
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 section 1.5, although the state position, velocity and acceleration are now 2D. The process error considers a maximum acceleration increment of 1 m/s^2, and the Unscented Transform supplies measurements of the target position 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 position and orientation estimate, and in the target detection in each image. This is clearly seen in the trajectories of Fig. 11 where the ground truth trajectory is a sequence of straight lines. These errors are accounted for by the Unscented Transform to estimate a covariance for the target observation, but nevertheless, the original target trajectory is filtered by a low pass filter before the input of the Kalman Filter. Analyzing the spectrum of the trajectory of the walking person, most of the energy is concentrated below 1 Hz. As the frequencies involved are too small, a low pass filter with too large attenuation or too small cut frequency would filter out true signal features such as going from zero velocity to motion in the beginning of the movement, and introduce delays in the filtered signal after curves. Therefore after empirical testing, the low pass filter parameters were set to a cut frequency of 2 Hz and attenuation of −10 dB. Thus the input of the Kalman Filter is a better conditioned signal, and the final trajectory is smoother.

3. Results

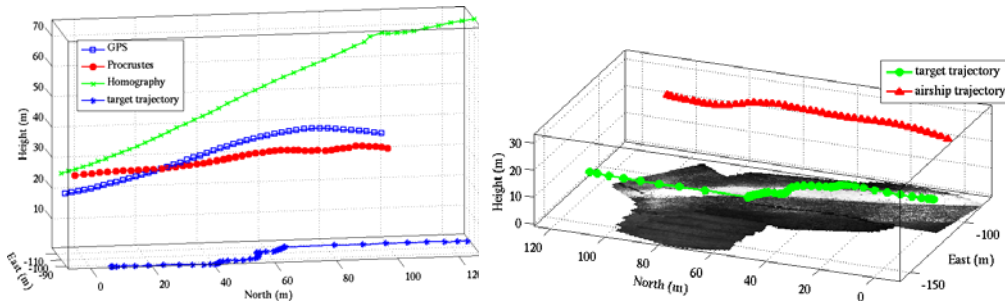
3.1 Tracking of a Moving Target from Airship Observations

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 in the image sequence 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.

Figure 5(a) shows the recovered airship trajectories using the method of section 1.4 (red circles) and by the standard homography estimation and decomposition method (green crosses). The blue squares show the GPS measured trajectory. In the ground the target (a moving car) trajectory derived from the airship trajectory recovered by our method is shown as blue stars.

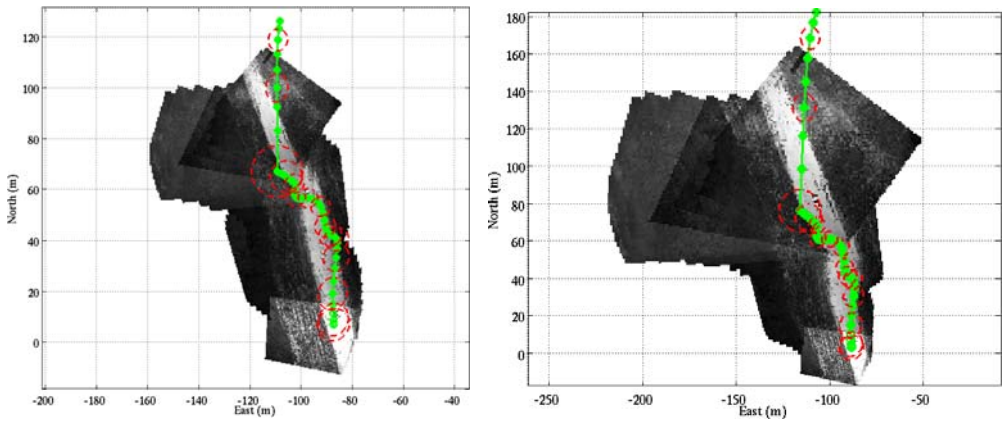
The trajectories recovered by the Procrustes method are shown again in figure 5(b). The images projected in the ground plane by using equation (2) to find the coordinates of their corners in the ground plane and drawing the image in the canvas accordingly. One every three images is drawn.

Figure 6 shows a 2D view of the target trajectory on the ground 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. The same low pass and Kalman filters were used with both methods.



(a) (b)

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.



(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 target trajectory with one standard deviation ellipses drawn in red.

3.2. Tracking after fusing GPS and Visual Odometry

In this experiment, the car has been driven in a closed loop in the ground while the airship was flying above it. To recover the airship trajectory, the translation recovered by the visual odometry was fused with GPS position fixes in a Kalman Filter with a constant acceleration model [Bar-Shalom et al., 2001]. The usage of this model does not imply that the actual acceleration of the vehicle or target is constant; it is just the approximation used by the filter. The GPS outputs standard deviation values for its position fixes (shown as the red ellipses and red points in Fig. 7), and the translation vectors from the visual odometry are interpreted as a velocity measurement between each pair of successive camera poses, with a manually set covariance smaller in the vertical axis than in the horizontal ones. The GPS absolute position fixes keep the estimated airship position from diverging, while the visual odometry measurements improve the trajectory locally. The fused airship trajectory is shown as green crosses in figure 7, while the target observations are shown as blue points in the ground plane. The target could not be continuously observed, therefore the straight lines (for example the straight lines crossing the path) indicate where observations were missing and resumed at some other point of the path.

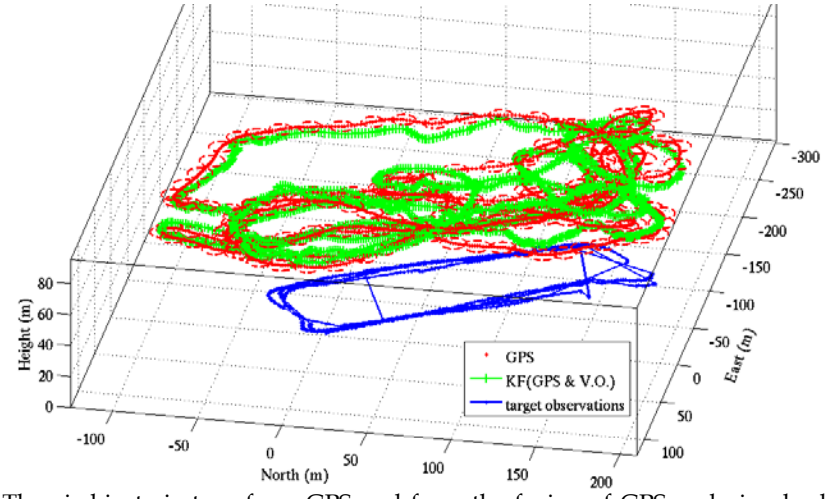


Fig. 7. The airship trajectory from GPS and from the fusion of GPS and visual odometry, with the target observations shown in the ground plane.

Figure 8(a) shows the target trajectory drawn over a satellite image of the flight area. The airship trajectory was taken directly from GPS. Figure 8(b) shows the same target trajectory obtained when the airship trajectory is recovered by a Kalman Filter fusing both visual odometry and GPS. In both figures, the squares show the coordinates of the target observations in the ground plane, the circles show the target trajectory filtered by its own Kalman Filter, and the crosses indicate that the target is "lost". The airship can not keep observing the target continuously, thus when there are not observations for an extended

period of time the tracked trajectory diverges. If the target position standard deviation becomes larger than 30 m than the target is declared "lost" and the filter is reinitialized at the next valid observation. Fusing the visual odometry with GPS resulted in a smoother trajectory for the tracked target.

3.3. 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 Procrustes method with the optimization described by equation (3) which improved the results (the AHRS was the less accurate MTB-9 model). The large squares in the floor provide a ground truth measure, as the person was asked to walk only on the lines between squares. The ground truth trajectories of the camera and the target person are highlighted in Fig. 9(a), and Fig. 9(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 measures 1.2 m.

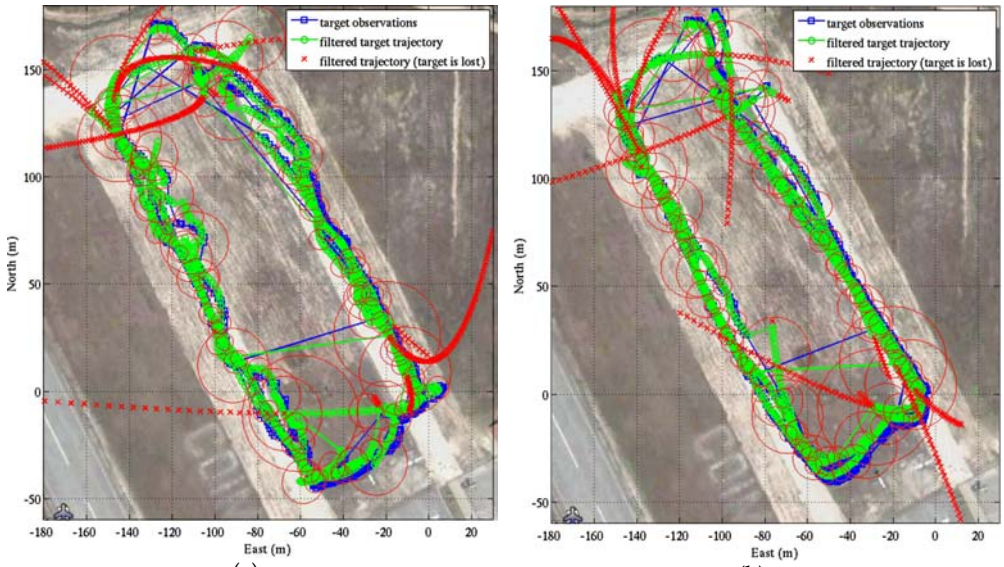
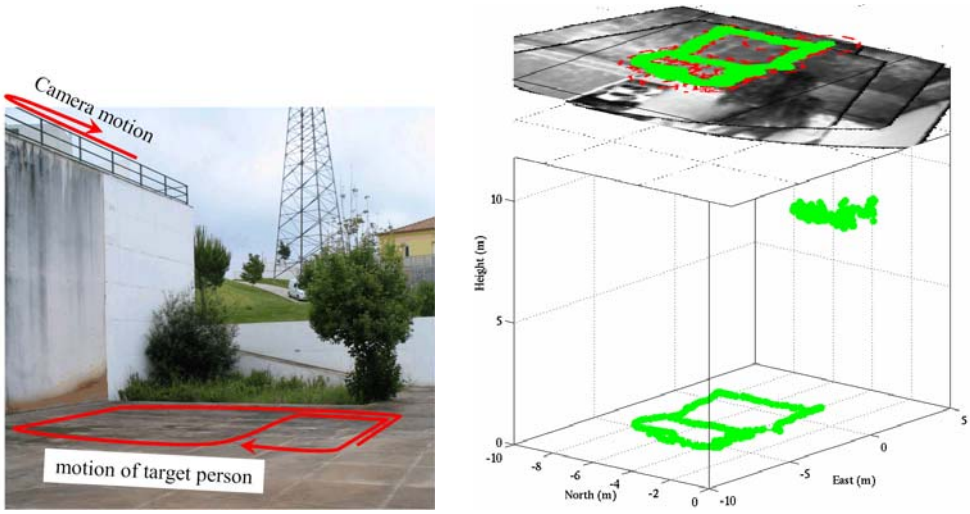


Fig. 8. The target trajectory over a satellite image of the flight area. The car followed the dirty roads. In (a), the airship trajectory was taken from GPS, in (b) a Kalman Filter estimated the airship trajectory by fusing GPS and visual odometry.

Figure 10 shows the target observations projected in the ground plane before (squares) and after (circles) applying a low pass filter to the data. Figure 11(a) shows a closer view of the target trajectory to be compared with Fig. 11(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 Kalman Filter. In both figures, the

large covariances in the bottom right of the image appear because the target was out of the camera field of view in a few frames, and therefore its estimated position covariance grew with the Kalman filter prediction stage. When the target comes back in the camera field of view the tracking resumed. The solid yellow lines are the known ground truth, marked directly over the floor square tiles in the image. 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. The same low pass filter and Kalman Filter were used to filter the target observations in both cases generating the target trajectories shown.



(a) (b)

Fig. 9. A photo with highlighted trajectories of camera and target person (a). A 3D view of the recovered trajectories, using the pure translation method to recover the camera trajectory (b).

4. Conclusions 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, and the filter parameters could be related to the camera and target motion characteristics.

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, especially in the height component, and thus improve the tracking performance.

Visual Odometry can also be fused with GPS position fixes in the airship scenario, and the improvements in the recovered airship trajectory translate in a smoother recovered trajectory for the moving target in the ground. As the GPS position fixes keep the system from diverging, the tracking can be performed over extended periods of time.
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. The visual odometry could also be fused with other sensors such as wheel odometry or beacon-based localization systems.

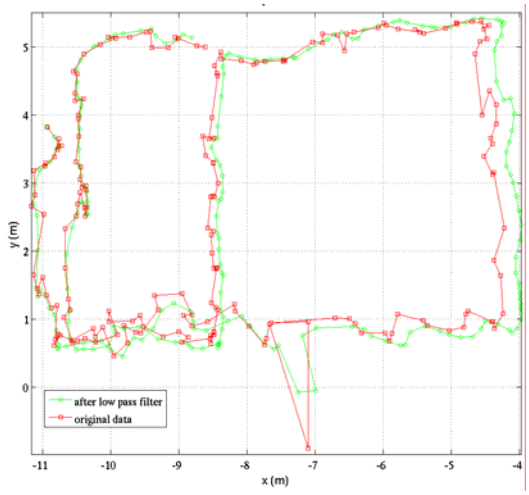
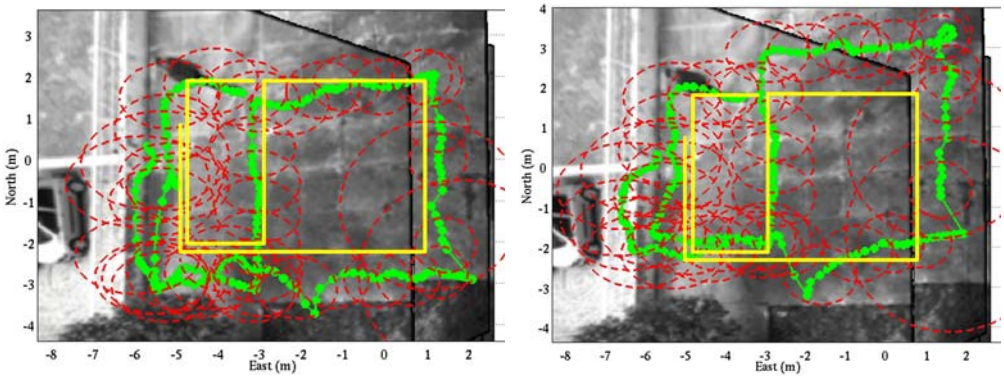


Fig. 10. A low pass filter is applied to the observed target trajectory before the input of the Kalman Filter.



(a) (b)

Fig. 11. The tracked trajectory of the target person. In (a) the camera trajectory was recovered by the pure translation method, in (b) by the image-only method.

5. References

Bar-Shalom, Y., Li, X. R., & Kirubarajan, T. (2001). *Estimation with Applications to Tracking and Navigation*. John Willey & Sons, Inc.

Bay, H., Tuytelaars, T., & van Gool, L. (2006). SURF: Speeded Up Robust Features. *In the Ninth European Conference on Computer Vision*, Graz, Austria.

Bouguet, J. (2006). *Camera Calibration Toolbox for Matlab*. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html.

Gower, J. C. & Dijksterhuis, G. B. (2004). *Procrustes Problems.* Oxford Statistical Science Series. Oxford University Press.

Hartley, R. & Zisserman, A. (2000). *Multiple View Geometry in Computer Vision.* Cambridge University Press, Cambridge, UK.

Julier, S. J. & Uhlmann, J. K. (1997). A new extension of the kalman filter to nonlinear systems. *In International Symposium in Aerospace/Defense Sensing, Simul. and Controls*, Orlando, FL, USA.

Lobo, J. & Dias, J. (2007). Relative pose calibration between visual and inertial sensors. *International Journal of Robotics Research*, 26(6):561□575.

Merino, L., Caballero, F., de Dios, J., & Ollero, A. (2005). Cooperative fire detection using unmanned aerial vehicles. *In IEEE International Conference on Robotics and Automation*, pages 1896□1901, Barcelona, Spain.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schafialitzky, F., Kadir, T., & van Gool, L. (2005). A comparison of affine region detectors. *International Journal of Computer Vision*, 65(7):43 - 72.

Mirisola, L. G. B. & Dias, J. (2007a). Exploiting inertial sensing in mosaicing and visual navigation. *In 6th IFAC Symposium on Intelligent Autonomous Vehicles,* Toulouse, France

Mirisola, L. G. B. & Dias, J. (2007b). Trajectory recovery and 3d mapping from rotation compensated imagery for an airship. *In IEEE Int. Conf. on Robots and Systems (IROS07)*, San Diego, CA, USA.

Mirisola, L. G. B. & Dias, J. (2008). Exploting inertial sensing in vision based navigation with an airship. *Journal of Field Robotics*. (submitted for publication).

Point Grey Inc. (2007). www.ptgrey.com.

Redding, J., McLain, T., Beard, R., & Taylor, C. (2006). Vision-based target localization from a fixed-wing miniature air vehicle. *American Control Conference, Minneapolis*, MN, USA.

XSens Tech. (2007). www.xsens.com.