Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Advanced 3D Localization by Fusing Measurements from GPS, Inertial and Vision Sensors

Fakhreddine Ababsa, *Member, IEEE*
IBISC Laboratory, CNRS FRE 3190
University of Evry-Val-d'Essonne
40, rue du Pelvoux, 91020 Evry, France
absbsa@iup.univ-evry.fr

***Abstract*— In this paper we present an efficient algorithm for estimating the 3D localization in an urban environments by fusing measurements from GPS receiver, inertial sensor and vision. Such hybrid sensor is important for numerous applications including outdoor mobile augmented reality and 3D robot localization. Our approach is based on non-linear filtering of these complementary sensors using a multi-rate Extended Kalman Filter. Our main contributions concern the modeling of the sensor fusion and the development of an efficient approach for camera pose tracking using only natural features. This method improves the accuracy of the estimated 3D localization. We evaluated the performances of our approach and demonstrated its effectiveness through experiments on real data.**

***Keywords*—augmented reality, sensor fusion, markerless tracking, extended kalman filter.**

I. INTRODUCTION

Accurate 3D localization in outdoor environment is crucial for mobile augmented reality systems [1]. Indeed, alignment of the real an virtual scenes requires knowing over the time and with high precision the camera pose in the world reference frame. This tracking problem has recently become a highly active area of research in augmented reality and computer vision. To deal with this problem, many approaches have been proposed in recent years, they can be divided into three main categories: pure vision-based, inertial and hybrid tracking solutions. Vision-based solutions [2][3] generally suffer from the high computational cost, sensitivity to noise and lack of robustness since, they depends on image feature extraction. In addition, since inertial sensors only measure the variation rate or accelerations, the output signals have to be integrated to obtain the position and orientation data. As a result, longer integrated time produces significant accumulated drift because of noise or bias [4]. Hybrid solutions attempt to overcome the drawbacks of any single sensing solution by combining the measurements of at least two sensors. The fusion of complementary sensors should be used to build better 3D localization systems. Synergies can be exploited to gain robustness, tracking speed and accuracy, and to reduce jitter and noise. State et al. [5] developed a hybrid tracking scheme which combined a fiducial–based vision and a magnetic tracker. They created a hybrid system that has the (static) registration accuracy of vision-based trackers and the robustness of magnetic trackers. Auer and Pinz [6], created a similar magnetic–vision system, in which their vision subsystem used corners as visual features. In their solution, prediction from the magnetic tracker is used to reduce the search areas in the optical tracking subsystem achieving a faster an more robust tracking. However, because of their low flexibility, magnetic trackers are not appropriate for mobile outdoor AR applications. Another popular choice in unprepared environments is inertial and natural feature video–based tracker fusion. You et al. [7] created a tracking system which combined a natural feature vision system with three gyro sensors. Their system limitation is that it just provide 3 DoF orientation tracking and not a full 6 DoF. The fusion approach is based on the SFM (structure from motion) algorithm, in which approximate feature motion is derived from the inertial data, and vision feature tracking corrects and refines these estimates in the image domain. Chen and Pinz [8] presented a structure and motion framework for real time tracking combining inertial sensors with a vision system based on natural features. Their model uses fusion data to predict user's pose and also to estimate a sparse model of the scene without any visual markers. An EKF is used to estimate motion by fusion of inertial and vision data, and a bank of separate filters to estimate the 3D structure of the scene. Chai et al. [9] employs an adaptive pose estimator with vision and inertial sensors for overcoming the problems of inertial sensor drift and vision sensor slow measurement. The extended Kalman filter (EKF) is used for data fusion and error compensation. Recently Bleser and Stricker [10] developed a markerless vision-inertial tracking system that works robustly in small-scale and large-scale environments, under varying lighting conditions and fast camera movements. However, their systems needs manual initialization and fails when vision data lacks. Ababsa and Mallem [11] developed a vision-inertial tracker which use a particle filter in order to fuse the sensor data. In [12] Diverdi and Hollerer introduced their GroundCam tracker, a vision-based method with high resolution and good short-time accuracy. They also demonstrated the feasibility of a hybrid tracker, coupling their GoundCam with a GPS receiver and a discrete beacon-based wide area sensor. In this paper, we present a novel hybrid approach for 3D localization in outdoor environments that integrates inertial, vision and GPS technologies. An original camera pose tracking using natural features forms the basis of our vision tracker. In order to fuse sensor data, we propose to use a multi-rate Extended Kalman filter.

The remainder of the paper is organized as follows. Section 2 is devoted to the hybrid tracker description. In Section 3 we explain the implementation of the EKF to solve the sensor

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

fusion problem for 3D localization. Experimental results are given in section 4. Conclusion and future work are presented in section 5.

II. HYBRID TRACKER OVERVIEW

Our 3D localization system combine an Inertial Measurement Unit (IMU) with a camera and a GPS receiver. The IMU is rigidly coupled with the camera and used to estimate the camera rotation (Figure 1). We used an Xsens MTi sensor which contains gyroscopes, accelerometers and magnetometers. The advantage of the MTi is that incorporates an internal digital signal processor which runs a real-time sensor fusion algorithm providing a reliable 3D orientation estimate. Data from MTi are synchronously measured at 100 Hz. For the vision, we opted for the uEye UI-2220-RE-C CCD camera with 6mm lens which is extremely compact, low-cost and well adapted for outdoor environments. Color images with resolution of 768×576 pixels at a frame rate of 25 Hz are streamed to a PC using a USB 2.0 connection. A Trimble GPS Pathfinder ProXT receiver mounted on a user provides GPS measurements. The ProXT receiver integrates a multipath rejection technology providing a submeter accuracy. Its rate update is about 1 Hz. The ProXT receiver uses a Bluetooth wireless connection, to communicate with the computer.

Figure 1. Sensor components of our Hybrid tracker

The three sensors providing measurements to the system are synchronized in hardware and runs at different rates. Furthermore, when working with heterogeneous sensors several coordinates systems are used. Each sensor provides data in its own reference frame. Indeed, the inertial sensor computes the orientation between a body reference frame R_I attached to itself and a local level reference frame R_G. Also, the GPS measurements are expressed according to the earth coordinates system. In the same way, the vision system estimates the camera pose with respect to the world coordinate system R_W. So, in order to fuse data from these different sensors, a transformation to a global fusion coordinate system is necessary. In our application, we need to estimate the 3D camera pose (position + orientation) according to the world coordinates system R_W. Thus, the orientation given by the IMU and the position measured by the GPS must be expressed in R_W. The transformations between the fusion coordinate system (R_W) and the local sensor coordinates are computing using two calibration processes : inertial/camera and GPS/camera. These calibration methods are described in detail in our early paper [13]

III. SENSOR FUSION FOR 3D LOCALIZATION

An Extended Kalman filter is used for fusing 2D/3D points correspondences from the image analysis, the orientation given by the inertial sensor and the position measured by the GPS receiver. This fusion allows to obtain an optimal camera pose estimate. In our approach the vector state X consists of the position and the orientation of the camera with respect to the world coordinate system R_W. For computational we use a unit quaternion to represent the rotation. Thus, the state vector is given by:

$$X = \begin{bmatrix} q_0 & q_x & q_y & q_z & t_x & t_y & t_z \end{bmatrix} \quad (1)$$

We denote the camera state at time t by the vector X_t. The EKF is used to maintain an estimate of the camera state X in the form of a probability distribution $P(X_t | X_{t-1}, Z_t)$, where Z_t is the measurement vector at time t. The equations of the EKF are well known [14][15] and are not given here. As the measurement data come from three heterogeneous sensors independent of each other, so each sensor will have its own measurement model. The filter will perform the time update step when either inertial, GPS or camera data is available. Then a measurement update step will be followed to update the filter's state according to the new measurements input.

A. Time and measuremenrt update

The time update model is employed in order to predict the 3D localization of the camera at the following time step. In our approach we don't incorporate the kinematics of the camera motion in the state vector, so the time update equation can be simply given by :

$$X_t^- = A \cdot X_{t-1} \quad (2)$$

Where A is 7×7 identity matrix.
The time update step also produces estimates of the error covariance matrix Σ from the previous time step to the current time step t. To perform this prediction we use the general update equation of the Kalman filter :

$$\Sigma_t^- = A \cdot \Sigma_{t-1} \cdot A' + Q_{t-1} \quad (3)$$

Where Q_t represents the covariance matrix of the process noise. Σ_t reflects the variance of the state distribution.

In the measurement update, the filter updates the state vector X_t according to the input measurement. The measurement update step is executed when either one of the three sources of data becomes available. For EKF, the prediction of the sensor output is given by:

$$z_t = h(X_t) + v_t \quad (4)$$

Each of the three sensors has its own output equation h and an uncertainty in the output space. So, we present the three different measurement models.

B. Camera measurement model

Our goal is to estimate the camera pose using only natural points. We consider a pin-hole camera model and we assume that the intrinsic camera parameters are known. The world coordinate frame is a reference frame. All the 3D model points are defined with respect to it. Formally, a 2D projection $m_i = [u_i \quad v_i \quad 1]^T$ of a 3D point $p_i = [x_i \quad y_i \quad z_i \quad 1]^T$ on the normalized image plane (figure 1), is expressed as (in homogenous coordinates) [16]:

$$s \cdot m_i = K \cdot [R \quad T] \cdot p_i \tag{5}$$

Where the matrix K contains camera calibration parameters, such as focal length, aspect ration and principal point coordinates and s is a scale factor. The 3×3 rotation matrix R and the translation vector T describe the rigid body transformation from the world coordinate system to the camera coordinate system and are precisely the components of the state vector X.

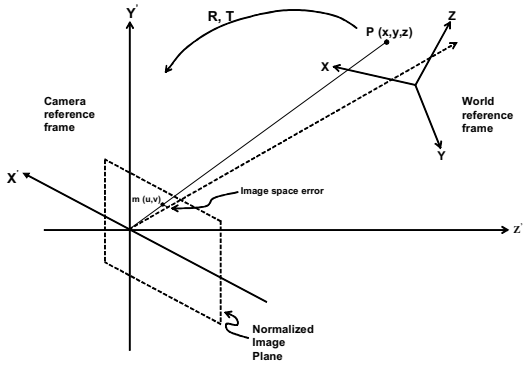

Figure 2. 2D/3D projection points model.

Since equation 5 gives a relationship between 3D model points, their corresponding 2D extracted image points and the camera pose parameters, then the camera measurement equation can be written, for each matching event $p_i \rightarrow m_i$:

$$z_t^{camera} = h_{camera}(X_t) + v_t^{camera} \tag{6}$$

Where

$$\begin{cases} z_t^{camera} = [u_i \quad v_i]^T \\ h_{camera}(X_t) = \left[\left(\frac{s \cdot m_i}{s} \right)_u \quad \left(\frac{s \cdot m_i}{s} \right)_v \right] \end{cases} \tag{7}$$

v_t^{camera} represents the noise term in the camera measurement input with covariance R_t^{camera} . The noise is due to the uncertainty in the measured image position of the extracted 2D points. The non linear function $h_{camera}(X)$ in measurement equation (6) relates the state to the camera measurement input. Three 2D-3D points correspondences are sufficient in theory to recover 6-DOF camera pose through in practice mores points may be required to increase the accuracy.

C. Inertial measurement model

The IMU sensor will produces the camera orientation with respect to the world coordinate system. This information will be associated with the rotation term in the state vector. The inertial measurement model is then given by:

$$\begin{aligned} z_t^{inertial} &= h_{inertial}(X_t) + v_t^{inertial} \\ &= H_{inertial} . X_t + v_t^{inertial} \end{aligned} \tag{7}$$

Where $H_{inertial} = [I_{4\times4} \quad 0_{3\times3}]$ and $v_t^{inertial}$ represents the noise term in the inertial measurement input with covariance $R_t^{inertial}$. The noise specification are given by the IMU constructor.

D. GPS measurement model

The used GPS receiver measures, with a high accuracy, the X-Y positions of the camera according to the world coordinate system. In addition, the GPS altitude accuracy is usually poorer than horizontal accuracy due to the system's basic technique of measuring, the error is about several meters. So, in our application, as the camera is handheld by the user, we fixed Z at the height of the camera according to the ground. The GPS is then defined as follows:

$$\begin{aligned} z_t^{GPS} &= h_{GPS}(X_t) + v_t^{GPS} \\ &= H_{GPS} . X_t + v_t^{GPS} \end{aligned} \tag{8}$$

Where $H_{GPS} = [0_{4\times4} \quad I_{3\times3}]$ and v_t^{GPS} represents the noise term in the GPS measurement input with covariance R_t^{GPS} . The noise specification are given by the GPS receiver constructor.

E. Fusion Algorithm

The goal of the fusion filtering is to estimate the 3D localization of the camera using a multi-rate information from the vision, the inertial and the GPS receiver sensors. One iteration of the fusion algorithm can be summarized as follows:

Algorithm	Recursive 3D Localization
1.	Perform an initialization of the state vector and the covariance matrix : X_0 and Σ_0
2.	Perform the time update step from t-1 to t using the process model (2) and (3)
3.	If input data is a new image from the camera, perform the camera measurement update using (9).
4.	If input data is inertial data, perform the inertial measurement update step using (11).

5. If input data is GPS receiver data, perform the GPS measurement update step using (12).
6. Set t=t+1 and iterate from step 2.

Our implementation runs in real-time with 100 Hz for inertial measurements, 25 Hz for the camera frame rate and 1 Hz for the GPS receiver. Also, we consider that measurement noises for the three sensors are Gaussian. The time update step is performed at each sampling period. If the three sets of data are sampled from the three sensors at the same time, the time update step is executed once. But the measurement update step is processed three times since the three data sets come from different sensors.

IV. EXPERIMENTS

We evaluated the performance of our 3D localization technique in a real outdoor environment. For that, a user explored the urban scene (figure 3-a) within the hybrid sensor (figure 3-b). The 3D model of the building is known, it is composed of 120 natural points defined by their 3D coordinates within the world coordinates frame. Several trials in different locations were recorded. The data coming from the three sensors are time-stamped and stored in data file.

(a) (b)

Figure 3. Experiments setup.

The camera pose estimated by the visual tracking was used as ground truth for the performance evaluation. We have defined a set of 2D/3D points correspondences from the first sequence frame. The interest points are then tracked frame to frame using the KLT feature tracking algorithm [17][18]. The 2D/3D correspondences are then updated and used to estimate the 3D localization. To investigate the effectiveness of the fusion algorithm to estimate the 3D localization when the vision-tracking fails, measurement image data, namely the extracted 2D points, are perturbed by Gaussian noise. Each 2D image point from the 2D/3D points correspondence set, used by the camera measurement model to estimate the camera pose, is perturbed by Gaussian noise with standard deviation of σ pixels. 3D localization is then estimated using the fusion algorithm and the errors between the truth data and the poses given by fusion are computed. Figure 4 show the X and Y average error values (mean and the standard deviation) for different noise levels. We can see that the mean error increases with increasing noise level. Also, for $\sigma > 3$, the mean errors given by the fusion filter are always lower than the ones given by the vision system when used alone. This demonstrates the effectiveness of our fusion algorithm to estimate the camera localisation in presence of image outliers. For example, for σ=8 pixels, the mean and the standard deviation given by our fusion filter are (0.80m,0.76m) for T_x and (0.77m,0.65m) for T_y. Figure 5 shows the recovered camera trajectory in world coordinate frame when using noised image features. We note that the fused trajectory of the three sensors is closer to the ground truth than the one estimated using only vision.

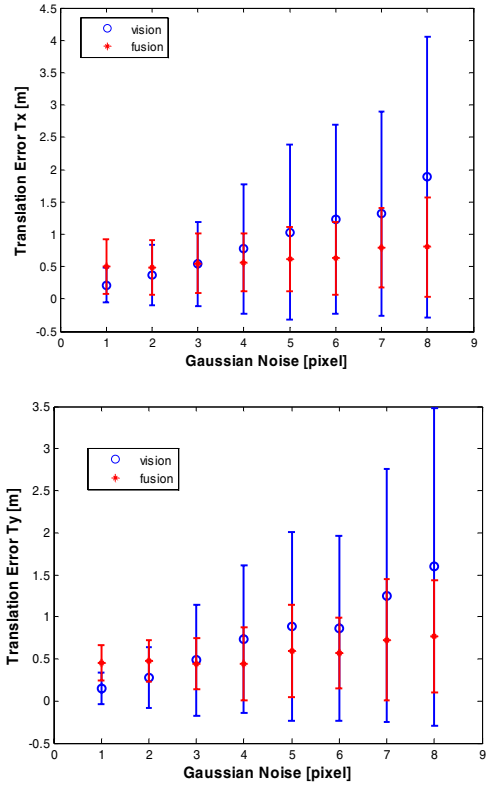

Figure 4. Position errors in presence of image outliers

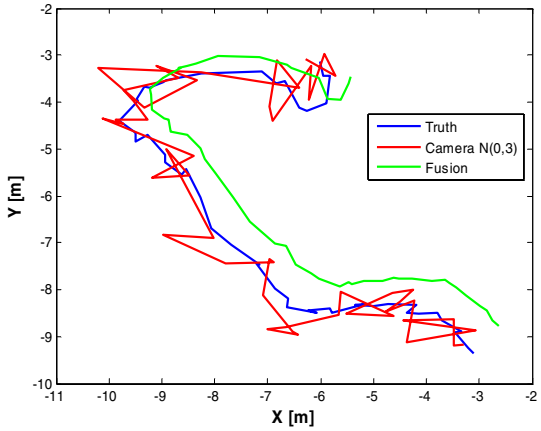

Figure 5. Camera Trajectory in world coordinate system

Finally, the results of our fusion algorithm are compared to the state-of-the-art results reported by Diverdi and Hollerer [11]

with their GroundCam system which also combines a camera, a GPS receiver and an orientation tracker. The authors run their system along a residential street for approximately 90 seconds and compare the estimated trajectory to a hand-labelled ground truth. The reported RMS is about 5.5m. We run our system in the similar conditions around our institution building. The obtained RMS is about 1.55m which shows that our fusion algorithm generates results significantly better.

V. CONCLUSION

We have presented a 3D localization system which allows to obtain real-time camera poses by fusing vision, inertial and GPS measurements using a multi rate EKF. Experiments are done in a real outdoor environment and demonstrated the robustness and the accuracy of our system. Indeed, the addition of GPS and IMU yield a robust system which can handle vision tracking failure. Future work aims to improve the robustness of our vision tracker by integrating robust natural features detection and matching techniques like SIFT [19] or SURF [20] methods.

REFERENCES

[1] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre: "Recent Advances in Augmented Reality". IEEE Computer Graphics and Applications, vol 21, no 6, pp. 34-47, 2001.

[2] R. I. Harrtley, and A. Zissermann, "Multiple View Geeometry in Computer Vision". Cambridge University Press, 2004.

[3] F. Ababsa, M. Maidi, J-Y. Didier, M. Mallem - Vision-Based Tracking for Mobile Augmented Reality". Multimedia Services in Intelligent Environments, Springer pp. 297-326, 2008

[4] P. Lang, A. Kusej A. Pinz and G. "Brasseur. Inertial tracking for mobile augmented reality". In Proc. of the IMTC 2002, Anchorage, USA. Vol. 2, pp. 1583-1587, , 2002.

[5] A. State, G. Hirota, D. T. Chen, W. F. Garrett, and M. A. Livingston, "Superior augmented reality registration by integrating landmark tracking and magnetic tracking," in SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, (New York, NY, USA), ACM Press, pp. 429–438, 1996.

[6] T. Auer and A. Pinz. Building a hybrid tracking system, "Integration of optical and magnetic tracking". In Proc. of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR '99), (Washington, DC, USA), pp. 13-19, 1999

[7] S. You, U. Neumann and R. Azuma. "Hybrid Inertial and Vision Tracking for Augmented Reality Registration". In Proc. of IEEE International Conference on Virtual Reality (VR'99), Lafayette, Louisiana, USA, pp. 260-267, 1999.

[8] J. Chen and A. Pinz. "Structure and motion by fusion of inertial and vision-based tracking". In Digital Imaging in Media and Education, Proc. of the 28th OAGM/AAPR Conference, pp. 55-62, . 2004.

[9] L. Chai, W. Hoff and T. Vincent. "Three-Dimensional Motion and Structure Estimation Using Inertial sensors and Computer Vision for Augmented Reality". Presence: Teleoperators & Virtual Environments, pp. 474-492, 2002.

[10] G. Bleser, and D. Stricker. "Advanced tracking through efficient image processing and visual-inertial sensor fusion". In Proc. of IEEE International Conference on Virtual Reality (VR 08), Reno, Nevada, USA, pp. 137-144, 2008.

[11] F. Ababsa, and M. Mallem. "Hybrid three-dimensional camera pose estimation using particle filter sensor fusion". Advanced Robotics, vol. 21, no. 1, pp. 165-181, 2007.

[12] S. DiVerdi, and T. Höllerer, "GroundCam: A Tracking Modality for Mobile Mixed Reality". In Proc. of IEEE International Conference on Virtual Reality (VR 07), Charlotte, North carolina, USA, pp. 75-82, 2007.

[13] I. M. Zendjebil, F. Ababsa, J.Y. Didier, and M. Mallem. "On the hybrid aid-localization for outdoor augmented reality applications". In Proc. Of the 15th ACM Symposium on Virtual Reality Software and Technology (VRST 08), Bordeaux, France, pp. 249-250, 2008.

[14] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". Transactions of the ASME, Journal of Basic Engineering, vol. 82, pp. 35-45, 1960.

[15] G. Welsh, and G. Bishop, "An introduction to the Kalman Filter". Technical report, Univrsity of North Carolina at Chapel Hill, Departement of computer science, 2001.

[16] O. Faugeras. "Three-Dimensional Computer Vision: A Geometric Viewpoint" MIT Press, Artificial Intelligence Collection, 1994.

[17] C. Tomasi and T. Kanade. "Detection and tracking of point features". Technical report, Carnegie Mellon University, 1991.

[18] J. Shi, and C. Tomasi. "Good features to track". In Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 94), Seattle, WA, USA, pp 593-600, 1994.

[19] D.G. Lowe. "Object recognition from local scale-invariant features". In Proc. of IEEE International Conference on Computer Vision (ICCV 99), Kerkyra, Corfu, Greece, vol. 2, pp. 1150–1157., 1999.

[20] H. Bay, T. Tuytelaars, and L. Van Gool. "SURF: Speeded Up Robust Features". In Proc. of the 9th European Conference on Computer Vision, Springer (ECCV 06), Graz, Austria, LNCS volume 3951, part 1, pp 404--417, 2006.