Seamless Indoor/Outdoor 6 DOF Tracking of Trainees and Weapons

Supun Samarasekera, Rakesh Kumar, T. Oskiper, Z. Zhu, H.P. Chiu, R. Hadsell, L. Wang
Sarnoff Corporation
Princeton, NJ
{ssamarasekera, rkumar, toskiper, zzhu, hchiu, rhadsell, lwang}@sarnoff.com

ABSTRACT

The military and security forces maintain multiple MOUT sites to train personnel for dangerous urban operations. Some of these facilities are instrumented for automatic tracking and recording of an individual trainee's actions; this is done to adapt the training conditions in real time and to provide detailed information for after action reviews. Today, tracking capabilities are quite limited, and generally include video cameras installed across the facility and GPS systems for outdoor tracking. No cost effective systems exist that are capable of tracking the location, pose, and gaze direction of individual trainee and the location and pose of their weapons both indoors and outdoors. There is a need for systems that can provide such measurements over wide areas, such as MOUT sites that cover multiple square miles and include numerous buildings.

In this paper we present a system for tracking the trainee's location, head orientation, and weapon orientation that provides high precision and does not require an instrumented site. Tracking is achieved only with sensors mounted on the individual trainees. These sensors include helmet-mounted video cameras and an inertial measurement unit. The vision system estimates both relative motion based on visual odometry and absolute position and orientation based on landmark matching. The 3D landmark database is built autonomously prior to the exercise. The system seamlessly handles transitions into and out of GPS-denied environments (buildings, dense forests) by maintaining pose relative to what the cameras are seeing in addition to GPS. We have demonstrated the viability of this technology in urban, rural, desert, forest and indoor environments on human-wearable platforms.

ABOUT THE AUTHORS

Mr. Supun Samarasekera is currently the Senior Technical Manager of the Mobile Vision Group at Sarnoff Corporation. He received his M.S. degree from University of Pennsylvania. Prior to joining Sarnoff, he was employed at Siemens Corp. Supun Samarasekera has 15+ years experience in building integrated multi-sensor systems for training, security & other applications. He has led programs for robotics, 3D modeling, training, visualization, aerial video surveillance, multi-sensor tracking and medical image processing applications. He has received numerous technical achievement awards for his technical work at Sarnoff.

Dr. Rakesh "Teddy" Kumar is currently the Senior Technical Director of the Vision and Robotics Laboratory at Sarnoff Corporation, Princeton, New Jersey. Prior to joining Sarnoff, he was employed at IBM. He received his Ph.D. in Computer Science from the University of Massachusetts at Amherst in 1992. His technical interests are in the areas of computer vision, computer graphics, image processing and multimedia. At Sarnoff, he has been directing and performing commercial and government R&D projects in the areas of training, robotics, video surveillance, 3D modeling and medical image analysis. He has been one of the principal founders from Sarnoff for multiple companies: VideoBrush, LifeClips and Pyramid Vision Technologies. Rakesh Kumar received the Sarnoff Presidents Award in 2009, and Sarnoff Technical Achievement awards in 1994 and 1996 for his work in registration of multi-sensor, multi-dimensional medical images and alignment of video to three dimensional scene models respectively. He was an Associate Editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence from 1999 to 2003. He has served in different capacities on a number of computer vision conferences and NSF review panels. He has co-authored more than 50 research publications and has received over 25 patents.

Seamless Indoor/Outdoor 6 DOF Tracking of Trainees and Weapons

Supun Samarasekera, Rakesh Kumar, T. Oskiper, Z. Zhu, H.P. Chiu, R. Hadsell, L. Wang
Sarnoff Corporation
Princeton, NJ
{ssamarasekera, rkumar, toskiper, zzhu, hchiu, rhadsell, lwang}@sarnoff.com

INTRODUCTION

As large-scale immersive training facilities have become more common, there is an increasing need for detailed, high-precision tracking of both soldier and weapon through complex environments. Currently, tracking is only available in instrumented training facilities or outdoors where GPS can be used, and high-precision tracking in 6 degrees of freedom of both trainee and weapon is not available at all. Without this tracking capability, accurate review and playback is missing and training is less effective. There is also a need for tracking in non-instrumented, dynamic facilities so that training sites can be set up quickly and adapted in real time. We present a user-worn system that provides real-time tracking of trainee and weapon, including orientation, in complex indoor and outdoor training environments. The tracking of trainees and weapon can be used for after action review for MOUT training, performance analysis and emerging applications such as Augmented Reality based training.

The tracking solution that we propose relies on synchronized sensor inputs from 4 calibrated cameras and an inertial measurement unit (IMU) mounted on the trainee's helmet (Figure 1). The cameras are placed in two stereo pairs, one forward-facing and one rear-facing, which allows for robust tracking even if one pair is completely occluded. Visual odometry provides relative estimates of position, as does the IMU, and an extended Kalman filter is used to fuse the measurements and give a stable position estimate with very low latency. In addition to these algorithmic components, absolute positioning in a common coordinate frame is needed so that multiple trainees can be tracked together. Absolute positioning is also needed to negate the inevitable drift that occurs in any navigation system that measures relative movements. GPS gives absolute positioning, but the accuracy can be quite low and it is not available in many environments. Instead we build a landmark database composed of 3D landmarks that are recognized and used by the trainee-worn systems to infer absolute position in a common coordinate frame. The database is efficiently cached for real-time access and is critical for seamless indoor and outdoor tracking.

We also propose an innovative approach for building the landmark database that ensures very high accuracy of the landmarks while simultaneously producing a 3D model of the training site, both indoors and out. Calibrated visual and 3D data is collected using a robot outfitted with LIDAR sensor and cameras mounted on a pan-tilt unit. The robot traverses the training facility, populating the landmark database with 3D landmarks and building an integrated point cloud of the entire site. Algorithms applied to align the LIDAR data are also used to correct the pose of the 3D landmarks, producing a unified model and database with very high accuracy.

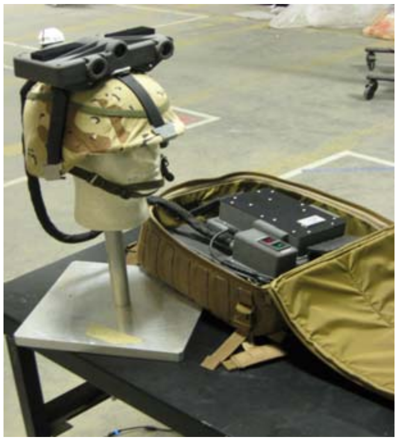


Figure 1: Helmet with forward and backward facing stereo cameras and MEMS IMU and ruggedized laptop processing units shown in back-pack.

The position and orientation of the weapon need to be estimated at the same level of accuracy as the trainee. To do this, fiducials are mounted on the weapon and tracked using the forward-facing helmet mounted cameras. This provides an accurate estimate of the weapon's position as long as it is in view of the helmet cameras, which is the case when it is being aimed or fired.

PREVIOUS WORK and OUR APPROACH

Current systems used for tracking trainees at a MOUT require significant infrastructure to be installed in advance. Time-consuming procedures are required to prepare the environment. There are very few systems which can track marines both indoors and outdoors. GPS-based systems [Saab'2010] may be used for providing location outdoors. However, the performance of these outdoor-only systems decreases in challenging GPS limited situations. UWB based systems have been used for indoor tracking of trainees to foot (30 cm) level accuracies [Fontana'2002] but do not provide orientation information. Finally none of these systems meet the challenging requirement for augmented reality where both location and orientation of the user's head must be tracked to centimeter level accuracy and within 0.05° accuracy for orientation. Overall, providing high accuracy tracking over large indoor and outdoor areas (multiple square miles) is a very challenging problem.

Real-time tracking by fusing visual and inertial sensors has been studied for many years with numerous applications in robotics, vehicle navigation and augmented reality. However, it is still unclear how to best combine the information from these complementary sensors. Since inertial sensors are suited for handling no or poor vision situations due to fast motion, occlusion, smoke, etc., many researchers use inertial data as backup [Aron'2007] or take only partial information (gyroscopes) from IMU [You'2001], [Reitmayr'2006], [Jiang'2004] to support vision-based tracking systems. To better exploit inertial data, several researchers use an extended Kalman filter to fuse all measurements uniformly to a pose estimate. They combine the filter with vision tracking techniques based on artificial markers [Foxlin'2003], feature points, or lines. These systems show that the vision measurements effectively reduce the errors accumulated from IMU. However, most of them conduct experiments on either synthetic data [Rehbinder'2003] or simulated vision measurements [Hol'2006]. Some systems provide results on realistic data, but within simple test environments [Schon'2007] or small rooms [Bleser'2009]. Moreover, they cannot eliminate the problem of long term drift over large areas inherent in inertial-based navigation platform. Due to recent advances in the image searching techniques, real-time landmark matching with a large landmark database has become possible [Nister'2006], [Se'2006]. Zhu et al. [Zhu'2008] integrated visual landmark matching to a pre-built landmark database in a visual-inertial navigation system. The continuously updating landmark matching corrects the long term drift in the system, and thus improves the overall performance. However, in that approach IMU data is mainly used for transitions between views where visual features are lost due to fast motion or bad illumination. Moreover, landmark matching, –whenever successful, is used to reset the pose solution in the global sense but it often lacks high-precision in pose estimation, which is required for augmented reality applications.

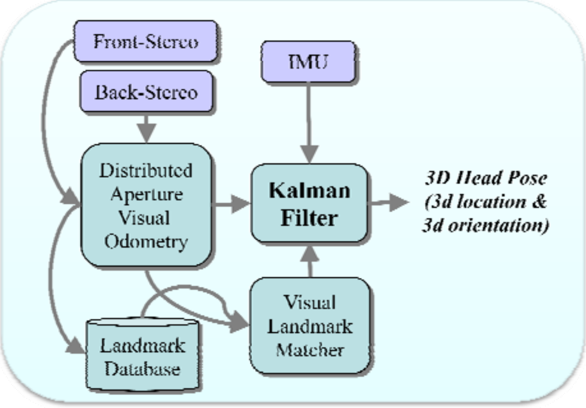

Front-Stereo
Back-Stereo
IMU
Distributed Aperture Visual Odometry
Kalman Filter
3D Head Pose (3d location & 3d orientation)
Visual Landmark Matcher
Landmark Database

Figure 2: Error-state Extended Kalman Filter block diagram with local and global external measurements.

There are two major differences between our work and other visual-inertial navigation systems. First, we adopt the *error-state* formulation [Roumeliotis'1999] in the extended Kalman filter. Under this representation, there is no need to specify an explicit dynamic motion model such as that used in [Oskiper'2007] for a given sensor platform. The filter dynamics follow from the IMU error propagation equations which evolve slowly over time and therefore are more amenable to linearization. The measurements to the filter consist of the differences between the inertial navigation solution as obtained by solving the IMU mechanization equations and the external source data, which in our case is the relative pose information provided by the visual odometry algorithm and global measurements provided by the visual landmark matching process (Figure 2). Hence, our Kalman filter framework incorporates two complementary vision measurements based on state-of-the-art vision tracking techniques. Relative pose measurements based on feature tracking between adjacent frames are usually located very precisely[Oskiper'2007]. Therefore, they do not jitter but suffer from drift or loss of track. Landmark matching [Zhu'2008] provides correspondences between fixed 3D features in a pre-built database and 2D points on the query frame. These measurements avoid drift but cause jitter. To make the outputted pose not only accurate but also stable, we fuse both local and global information in the extended Kalman filter.

In the remainder of the paper, we first present the 5 components of the system: [1] relative pose estimation through multi-camera visual odometry, [2] absolute pose estimation from visual landmark matching, [3] extended Kalman filter model for stable navigation, [4] simultaneous landmark collection and 3d model construction and [5] weapon pose estimation. Finally, we present experimental results, conclusions, acknowledgements and references.

RELATIVE MOTION ESTIMATION THROUGH MULTI-CAMERA VISUAL ODOMETRY

Visual odometry addresses the problem of estimating camera poses based on image sequences in a relative coordinate system. The poses of video frames are computed in the coordinate system of the first frame or a key frame in the sequence.

Video frames captured from the multi-camera system are used to compute the visual odometry solution. After acquiring the left and right camera image frames at time t_k, the first step consists of detecting and matching Harris corner based feature points in each stereo pair [Oskiper'2007]. Feature point image coordinates are normalized using the known intrinsic calibration parameters in each camera (by multiplication with the inverse of the calibration matrix) and compensated for radial distortion. In the stereo matching process, calibration information allows us to eliminate most of the false matches by applying epipolar and disparity constraints. This is followed by computation of the 3D locations corresponding to these feature points through stereo triangulation in the coordinate frame of the current left camera. Next, using the new image frames at time step t_{k+1}, 2D-2D correspondences are established by matching feature points between the previous frames at time step t_k and the current ones at t_{k+1}. This allows 3D-2D point correspondences to be established based on the 3D point cloud computed in the previous step. Finally, the pose of the left camera in each stereo pair can be computed using a robust resection method based on RANSAC followed by iterative refinement of the winning hypothesis where Cauchy-based robust cost function of the reprojection errors in both the left and right images is minimized. For the front stereo pair (j=1) and back stereo pair (j=2), this cost function is given by:

$$c_j\left(P_k^j\right)=\sum_{i=1}^{K_j}\rho(x_i^{l_j}-h(P_k^j X_i^j))+\rho(x_i^{r_j}-h(P^{s_j}P_k^j X_i^j))$$

where, for the j^{th} stereo pair, K_j is the number of feature points, x_i^l and x_i^r denote coordinates of the feature point i in the left and right images, X_i^j denotes its 3D position in homogeneous coordinates, P^s denotes the pose of the right camera in the left camera coordinate frame (known through stereo calibration), function h is used in denoting the conversion from homogeneous to inhomogeneous coordinates, $\rho(y) = \log(1 + ||y|^2/a^2)$ is the Cauchy-based robust cost function with a given scale parameter a, and finally $P_k^j=P_j(t_k,t_{k+1})$ is the relative pose across two time instants.

In our baseline approach all the above steps are performed independently for both the front and back stereo pairs in a parallel fashion. At the end of this process, two pose estimates are obtained from both pairs and the best one is chosen based on a cumulative (global) score [Oskiper'2007].

ABSOLUTE POSE ESTIMATION FROM 3D LANDMARK MATCHING

Landmark matching of the helmet camera videos to a landmark database allows us to locate and track the trainee's 3D position and pose in an absolute world coordinate system [Zhu'2008]. An incremental motion based navigation system fusing visual odometry, and IMU via Kalman filter is only locally accurate and will drift eventually as the errors accumulate. Landmark-based navigation locates the trainee in an absolute coordinate system and prevents drift. However landmark matching may not always succeed and also a solution based on it will jitter. Landmark matching is also computationally expensive and is not available for every frame in the video. Therefore, by integrating these two complementary modules (visual odometry and landmark matching), the two-stage localization technique dramatically increases the robustness of the combined system.

In our system, we define a landmark as a feature point in the scene. Specifically, it is extracted from the image using a Harris corner detector. Each landmark is associated with three elements: a 3D coordinate vector representing the 3D location, a 2D coordinate vector representing the 2D location in the image and a feature descriptor that characterizes the appearance. The histogram of oriented gradients (HOG) descriptor is used to model the appearance of each of the selected corner points.

The database is represented as a collection of landmark shots, where a landmark shot is a set of landmarks captured at a specific camera location and view point (or camera pose). A landmark shot is the basic unit of landmark matching. For each landmark shot, besides storing all the location (2D+3D) and appearance (HOG) information of each landmark into the database, the associated camera pose at that instant is also stored. We describe how we build a 3D model and landmark database in a following section.

When the system initializes, it locates itself by searching the whole landmark database. This is done via the fast indexing technique using a vocabulary tree

[Nister'2006]. Once the navigation system locates itself globally, it will update the current camera pose and its uncertainty to estimate a search region. The estimated search region will serve as a geo-spatial constraint to select a smaller set of landmarks for matching in the next frame. As a result, both efficiency and accuracy can be increased. During the mission, whenever the system fails to locate via landmark-based localization, the visual odometry and Kalman filter system takes over. The visual odometry and Kalman filter system localizes by estimating the frame-to-frame and IMU relative poses and integrating them over time. The system will resume landmark-based localization as soon as an image is found and matched in the landmark database.

EXTENDED KALMAN FILTER MODEL FOR STABLE NAVIGATION

We introduce a new Kalman filter framework to fuse IMU data, the local measurements from the distributed aperture visual odometry algorithm with front and back facing stereo cameras, and the global measurements from the visual landmark-matching module. Our Kalman filter adopts the so called "error-state" formulation, so there is no need to specify an explicit dynamic motion model such as the constant velocity process model used in our previous work. The filter dynamics follow from the IMU error propagation equations which vary smoothly and therefore are more amenable to linearization. The measurements to the filter consist of the differences between the inertial navigation solution as obtained by solving the IMU mechanization equations and the external source data, which in our case is the relative pose information provided by visual odometry algorithm and global measurements provided by the visual landmark matching process.

In our filter, we denote the ground (global coordinate frame) to camera pose as $P_{GC} = [R_{GC}\ T_{GC}]$ such that a point X_G in the ground frame can be transferred to the camera coordinates by $X_C = R_{GC}X_G + T_{GC}$. The total (full) states of our filter consist of the camera location T_{CG}, the gyroscope bias vector b_g, velocity vector v in global coordinate frame, accelerometer bias vector b_a and ground to camera orientation q_{GC}, expressed in terms of the quaternion representation for rotation:

$$\mathbf{s} = [\mathbf{q}_{GC}^T \quad \mathbf{b}_g^T \quad \mathbf{v}^T \quad \mathbf{b}_a^T \quad \mathbf{T}_{CG}^T]^T \ .$$

The state estimate propagation is obtained by the IMU mechanization equations with the gyroscope $\omega_m(t)$ and accelerometer $a_m(t)$ readings from the IMU between consecutive video frame time instants.

$$\dot{\hat{\mathbf{q}}}_{GC}(t) = \frac{1}{2}(\hat{\mathbf{q}}_{GC}(t) \otimes \hat{\boldsymbol{\omega}}(t))$$
$$\dot{\hat{\mathbf{v}}}(t) = \hat{\mathbf{R}}_{GC}^T(t)\hat{\alpha}(t) + \mathbf{g},$$
$$\dot{\hat{\mathbf{x}}}(t) = \hat{\mathbf{v}}(t), \qquad \dot{\hat{\mathbf{b}}}_g(t) = 0, \qquad \dot{\hat{\mathbf{b}}}_a(t) = 0$$

where

$$\hat{\omega}(t) = \omega_m(t) - \hat{\mathbf{b}}_a(t), \ \hat{\alpha}(t) = \mathbf{a}_m(t) - \hat{\mathbf{b}}_a(t)$$

and $\otimes$ is used to denote the quaternion product operation. The Kalman filter error state consists of

$$\delta\mathbf{s} = [\delta\mathbf{\Theta}^T \quad \delta\mathbf{b}_g^T \quad \delta\mathbf{v}^T \quad \delta\mathbf{b}_a^T \quad \delta\mathbf{T}_{CG}^T]^T$$

according to the following relation between the total state and its inertial estimate

$$\mathbf{q}_{GC} = \hat{\mathbf{q}}_{GC} \otimes \delta\mathbf{q}_{GC}, \ \delta\mathbf{q}_{GC} \simeq [1 \quad \delta\mathbf{\Theta}^T/2]^T$$
$$\mathbf{b}_g(t) = \hat{\mathbf{b}}_g(t) + \delta\mathbf{b}_g(t), \ \mathbf{b}_a(t) = \hat{\mathbf{b}}_a(t) + \delta\mathbf{b}_a(t)$$
$$\mathbf{v}(t) = \hat{\mathbf{v}}(t) + \delta\mathbf{v}(t), \ \mathbf{T}_{CG}(t) = \hat{\mathbf{T}}_{CG}(t) + \delta\mathbf{T}_{CG}(t)$$

The updating (correction) of the Kalman filter comes from two external sources of data: the relative pose information provided by the visual odometry algorithm and global measurements provided by the visual landmark matching process. To incorporate visual odometry poses that are relative in nature, we apply a stochastic cloning approach for our measurement model. In particular, these measurements are a function of the propagated error-state δs_2 and the cloned error-state δs_1 from the previous time instance, which require modifications to the original Kalman filter update equations [Roumeiotis,2002].

As for landmark matching, given a query image, landmark matching module returns the found landmark shot from the database establishing the 2D to 3D point correspondences between the query image features and the 3D local point cloud, as well as the camera pose belonging to that shot. First, every 3D local landmark point is transferred to the global coordinate system. After this transformation, the projective camera measurement model is employed so that each 3D point can be expressed in the current camera coordinate system. Then we apply the measurement equation in the error states of our filter to all the point correspondences returned as a result of landmark matching.

The Kalman filter fuses all the measurement data and allows for better handling of the uncertainty propagation through the whole system. In previous approaches [Zhu'2008] the Kalman filter output was used to locally propagate the navigation solution from one landmark match instance to another. In that case, the pose solution obtained as a result of landmark matching would effectively reset the filter output. By fusing both inertial and vision measurements, our system is also more robust under challenging conditions where there are insufficient visual clues on which to rely.

The other advantage of our Kalman filter is to eliminate a trade-off problem: landmark matching between the pre-built database and the current query frame provides global fixes to prevent the estimated poses from drifting during online tracking, but often lacks precision which results in jitter. We found that the accuracy of pose estimation from the landmark matcher decreases if there are few landmark point matches closer to the camera where the depth estimation is more accurate. To reduce the jitter, we capture the 3D local reconstruction uncertainty of each landmark point as a covariance matrix and implicitly rely more on closer points as global measurements in the Kalman filter. This provides more accurate and stable pose estimation.

SIMULTANEOUS LANDMARK COLLECTION AND 3D MODEL CONSTRUCTION

It is essential that the landmark database contain very accurate landmark positions, since matches to the database are used to correct the relative motion measurements given by visual odometry. Therefore, the landmarks are collected using a mobile robot outfitted with LIDAR sensor and cameras on a pan-tilt unit (see Figure 3). Because the 3D measurements from a LIDAR sensor have linear error with respect to range (unlike stereo 3D, which has quadratic error with respect to range), overlapping point clouds can be aligned with a high degree of accuracy. This provides the correction necessary to eliminate drift that would otherwise exist in the collected landmark database. The data is processed automatically, producing a point cloud model of the training site as well as a pose-corrected landmark database.

As the robot traverses the training site (autonomously or user-controlled), it stops at regular intervals and pans, recording full omni-directional visual and 3D data at each position. Each of these local data collections is called a 360° scan. The algorithm for automatic point cloud integration and pose correction is set up as a pipeline that processes each 360° scan in turn. A single 360° scan $\mathbf{S}_i$ includes LIDAR and camera data from time t_i to t_j. The LIDAR data consists of a set of scanlines $\mathbf{L}[t_i..t_j]$, the camera data consists of a set of images $\mathbf{I}[t_i..t_j]$, and the visual odometry outputs a corresponding set of poses (one for each image) $\mathbf{P}[t_i..t_j]$. There is a point cloud database $\mathbf{DB}_{3D}$ and a landmark database $\mathbf{DB}_{LM}$. Iterative closest point (ICP) is an EM-style of algorithm used to align two overlapping point clouds. The algorithm for constructing the model and landmark database follows:

```
for each scan $\mathbf{S}_i$
    integrate $\mathbf{L}[t_i..t_j]$ using $\mathbf{P}[t_i..t_j]$ to get $\mathbf{X}_i$
    query $\mathbf{DB}_{3D}$ for overlapping scans $\mathbf{X}_{DB}$
    align $\mathbf{X}_i$ with $\mathbf{X}_{DB}$ using ICP algorithm
    transform $\mathbf{X}_i$ with ICP correction:
                $\mathbf{X}_i' = P_{ICP} \mathbf{X}_i$
    add $\mathbf{X}_i'$ to $\mathbf{DB}_{3D}$
    transform $\mathbf{P}[t_i..t_j]$ with same ICP correction
                $\mathbf{P'}[t_i..t_j] = P_{ICP} \mathbf{P}[t_i..t_j]$
    add ($\mathbf{P'}[t_i..t_j]$, $\mathbf{I}[t_i..t_j]$) to $\mathbf{DB}_{LM}$
end
```

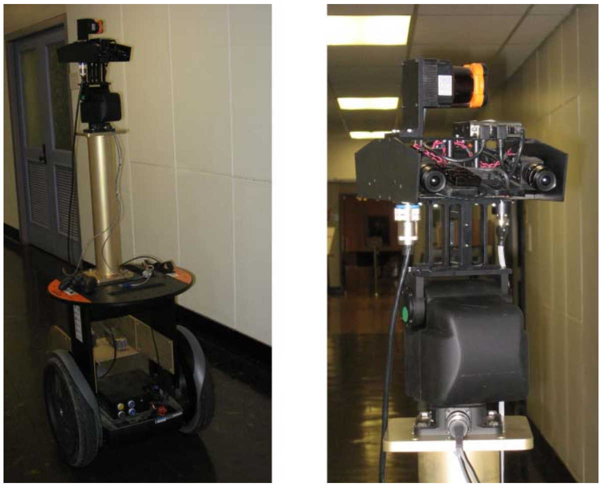


Figure 3: A segway robotic platform (RMP400) with LIDAR-camera sensor head. The sensor head is mounted at human height to record images from the same perspective as the trainee.

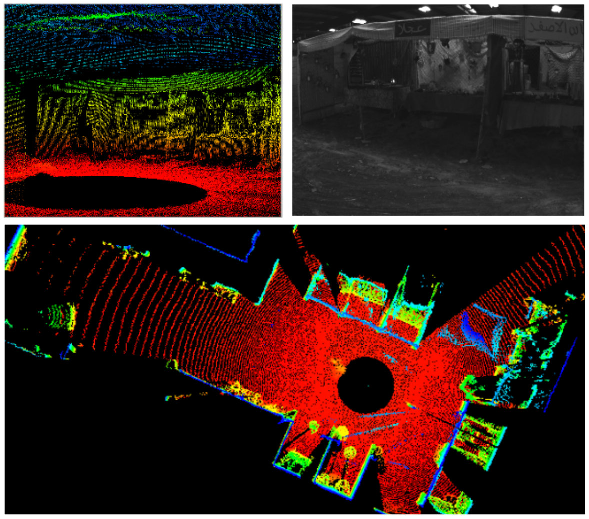


Figure 4 (Top left) View of a single point cloud in marketplace at IIT. (Top right) Camera view of same area of marketplace. (Bottom) Top down view of single marketplace scan.

Once all scans have been processed and the two databases have been populated with landmarks and point clouds, post-processing can be done to apply global transformations, remove redundant data, or subsample the point clouds to a uniform density. Figures 4 and 5 show examples of point clouds taken at the Immersive Infantry Trainer (IIT) at Camp Pendleton.

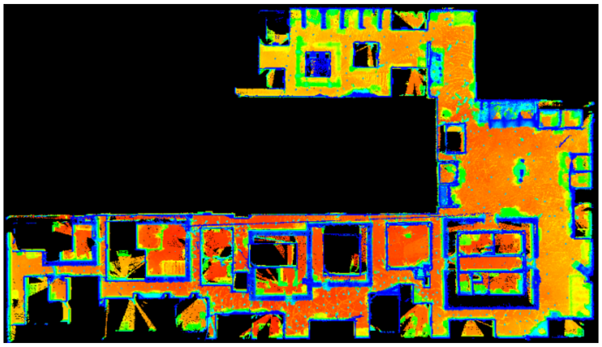


Figure 5 Top down view of entire IIT point cloud model (487 merged scans, subsampled to uniform 5 cm density).

WEAPON POSE ESTIMATION

As shown in Figure 6, multiple markers are mounted on a weapon. By detecting them in the video frames from the two forward-facing cameras, the 6-DOF pose of the weapon can be accurately estimated.

Fiducial detection is well-studied in augmented reality [Kato'1999, Claus'2004]. However, existing approaches are not reliable enough to handle the difficulties inherent in our application, including significant lighting changes, severe foreshortening, and the limited fiducial size. Different from most existing approaches that are based on adaptive thresholding, our approach largely relies on curve extraction. This leads to its robustness under non-uniform lighting conditions and its computational efficiency. It can run at 60Hz with 640x480 images on a single-core CPU.

The algorithm can be outlined as follows: Edge pixels are detected with non-maximum suppression on gradient and linked into curves. Rectangles are detected from the curves. Each rectangle is classified into one of the fiducials or non-fiducial based on the configuration of its inside curves. Each rectangle is also classified according to its inside black blobs, which are detected with thresholding. The more confident classification result is output.

Fiducials are detected in both of the left and right forward-facing cameras. Based on the correspondences between the detected 2D points of the fiducials' corners and their 3D points, the generalized 3-point algorithm in [Nister'2007] is used to calculate the pose of the weapon.

To test the accuracy of the estimated pose, a laser pointer is installed on the weapon. The 3D position and orientation of the laser beam relative to the fiducials are calibrated. According to the estimated weapon pose, we can draw the virtual projection of the laser beam in the images, such as the green line shown in Figure 6. If the pose is accurate, the image point of the highlighted spot where the laser beam hits an object (the bright spot on the wall in Figure 4) should be very close to the projection of the laser beam. In our experiments, the average distance is around 1 pixel in all images.

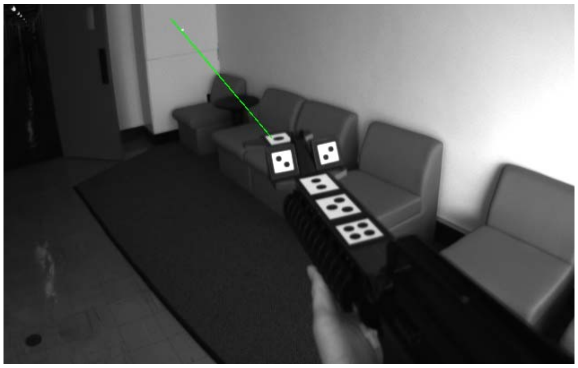


Figure 6: A panel with 6 markers, and a laser pointer are mounted on the gun. The green line is the estimated virtual projection of the laser beam. The bright spot on The upper-left wall is the highlighted laser point, and it is almost on the line in the image.

EXPERIMENTAL RESULTS

In this section, we report a number of experiments aimed at evaluating different aspects of the performance of our tracking framework. We also demonstrate that our framework can provide highly accurate real-time tracking in both indoors and outdoors over large areas. Compared to [Oskiper'2007] and [Zhu'2008], we show our navigation system can provide more stable pose estimation to fulfill the demanding requirements for augmented reality applications.

B. Performance of fusing local and global measurements

Our Kalman filter allows for better handling of the uncertainty propagation through the whole system, and is able to incorporate the global measurements which are 3D to 2D feature point correspondences from landmark matching [Zhu'2008]. To demonstrate the influence to our filter by incorporating these global measurements, we collected an outdoor sequence while the user wearing our system traveled along a predefined path. We constructed the landmark database of the area where the user would travel before-hand. The user traveled around 256 meters (3 minutes and 40 seconds long) and went back to the starting position. The results (Figure 7) shows that fusing global measurements reduces the 3D loop closure error estimated by our filter from 2.4873 meters (relative pose estimation alone) to 0.5712 meters (absolute pose estimation using both local and global measures).

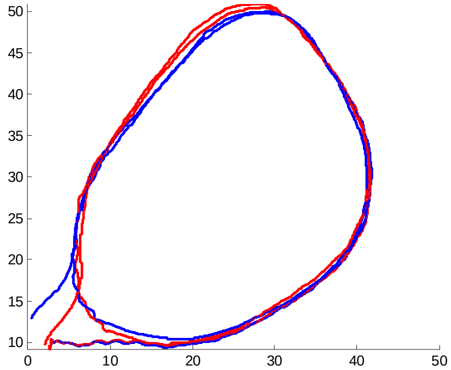


Figure 7: Blue line shows the estimated 3D trajectory by fusing IMU data and local measurements (relative pose by visual odometry). Red line showsthe estimated 3D trajectory by fusing IMU data, local measurements, and global measurements from landmark matching. The total traveled distance is around 256 meters.

C. Real-time tracking over large areas

To demonstrate that our system can be used in both indoors and outdoors over large areas, Figure 8 shows the automatically generated real-time camera trajectory corresponding to an 810 meter course within our campus completed by a user wearing our helmet, backpack system, and a video see- through HMD. This user walked indoors and outdoors in several loops. The entire area shown in the map is within the pre-built landmark database capture range which is loaded in the beginning before the exercise takes place. Landmark matches occur whenever a query image is within close proximity to a stored landmark shot in the database.

Figure 9 shows several screen shots corresponding to locations towards the beginning, middle and end of this exercise, obtained from our visualization tool which is used to verify the accuracy of the localization. This visualization tool uses the camera poses that are output by the system to render views from a 3D graphical model built upon the same visual data as the landmark database. We compare the rendered views to the actual video images. It is observed that these views are in very good agreement which indicate how precisely the camera is tracked throughout the entire duration of the course.

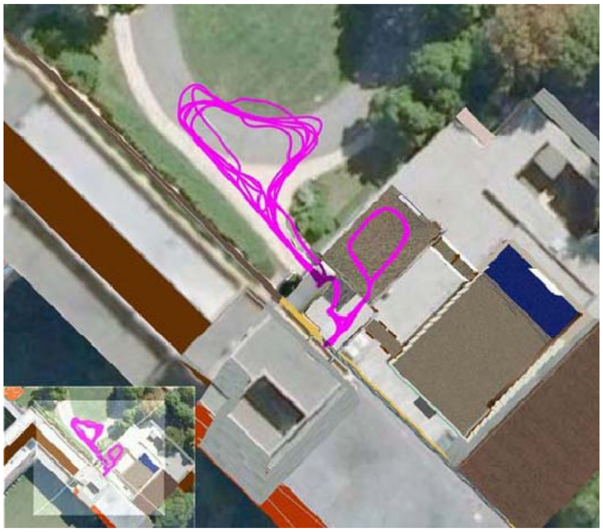


Figure 8: Real-time computed camera trajectory corresponding to an 810 meter path completed in 16.4 minutes during a live exercise.

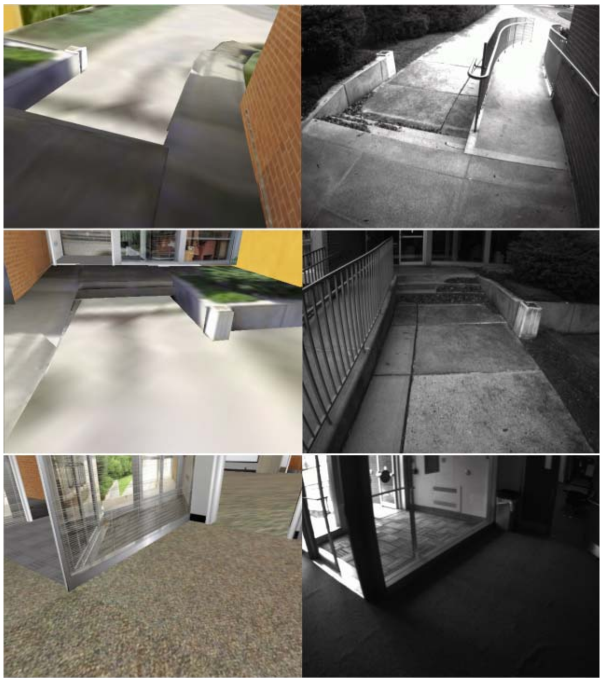


Figure 9: The views rendered from the model using the real-time camera pose estimates by our system for various locations throughout the exercise, together with the real scene views captured by the camera.

Figure 10 shows the automatically generated real-time camera trajectory corresponding to a 253.82 meter course within our campus completed by a user wearing our helmet and backpack system. The landmark database used for this example is described in Figures 4 and 5. This sequence was taken at the Immersive Infantry Trainer (IIT) at Camp Pendleton.

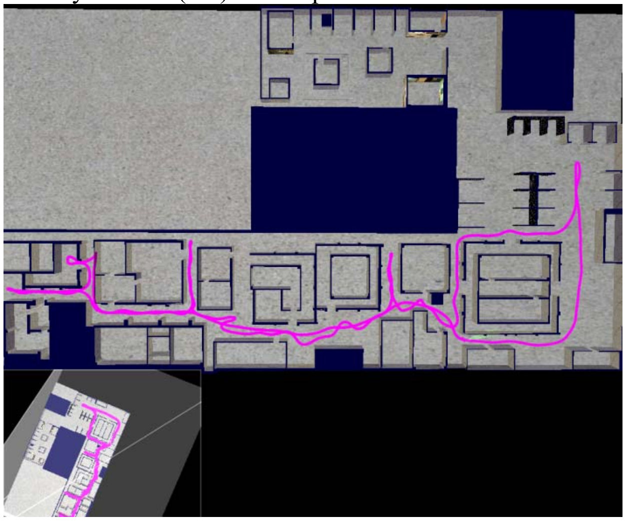
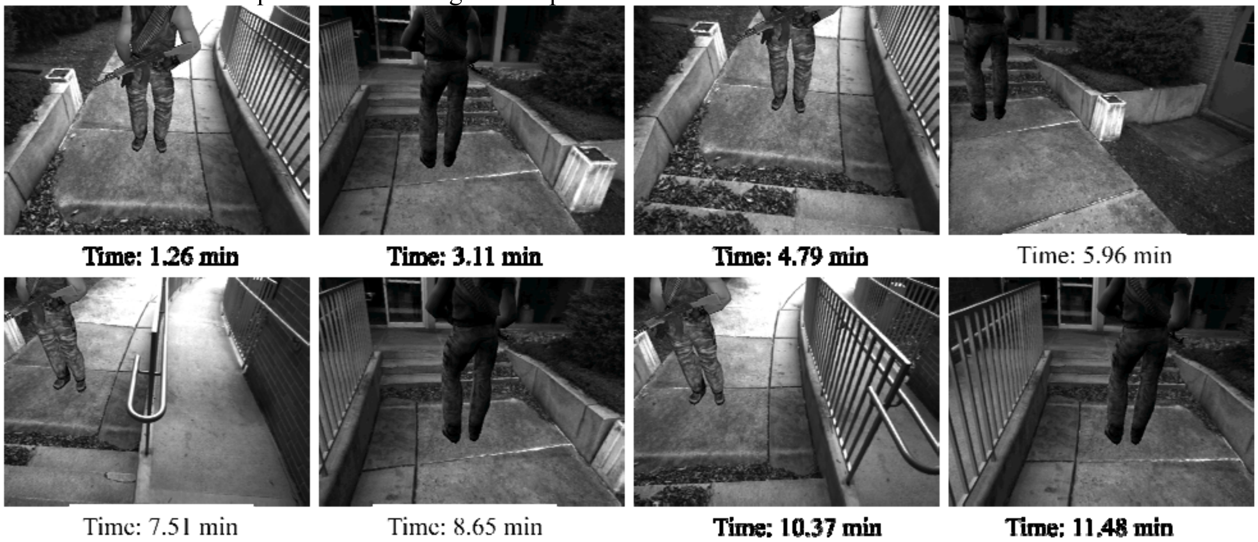


Figure 10: Real -time computed camera trajectory corresponding to a 253.82 m run over 5.64 minutes

D. Pose estimation for augmented reality

During the same 810 meter course described above, we inserted virtual actors at particular locations based on the estimated pose and recorded the insertion video which was seen by the user from the video-see-through HMD. For example, we inserted one virtual actor right outside the stone steps of a building. The pose estimation from our system needs to be very accurate and stable during the whole course, otherwise it will break the illusion of augmented reality for the user. Figure 11 shows 8 snapshots of the video when the user went through the entrance of the building at different times during the 16.4-minute course. The positions of the inserted actor are very consistent in these 8 snapshots. This result demonstrates that the system is able to provide stable, drift-free pose estimation for a long period, compared to our previous approach in [Zhu'2008].. Figure 12 shows the frame-to-frame pose translation estimated by [Zhu'2008] and our filter respectively. To save space, we only show the translation over a 450-frame period taken from the 16.4 minute video. Since the walking speed of the user doesn't change much in a very short period (such as one frame, 0.0677 seconds), the translation between frames should be very smooth. However, in [Zhu'2008], landmark matching disturbs the consistency of pose estimation due to the lack of high-precision. The peaks of the green curve in Figure 9 correspond to the jitter of inserted virtual actors viewed by the user. By capturing the 3D reconstruction uncertainty of landmark points and thus relying more on closer points as global measurements in the Kalman filter, our navigation system can reduce the jitter in pose estimation for augmented reality applications.



Time: 1.26 min

Time: 3.11 min

Time: 4.79 min

Time: 5.96 min

Time: 7.51 min

Time: 8.65 min

Time: 10.37 min

Time: 11.48 min

Figure 11: 8 snapshots taken from the video when the user went through the entrance of our building at different loops

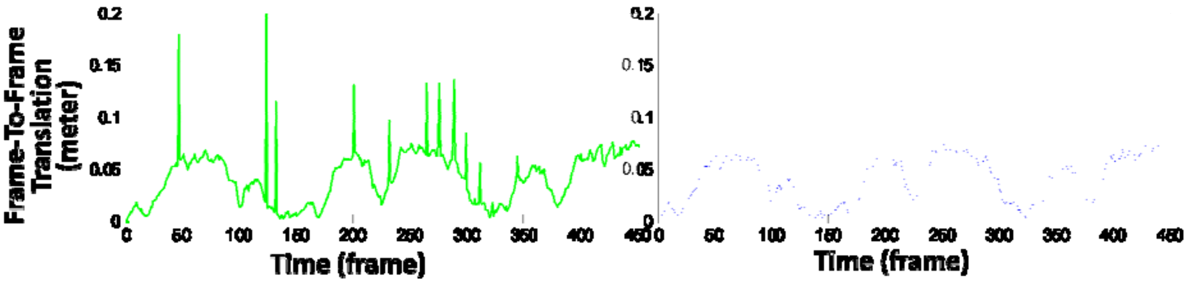


Figure 12: The frame-to-frame estimated translation computed by [Zhu'2008] (green) and our system (blue).

CONCLUSIONS

We presented an infrastructure-free 6 DOF tracking system based on a unified Kalman filter framework using local and global sensor data fusion for vision aided navigation related to augmented reality applications. We showed results to illustrate the accuracy and robustness of our system in both indoors and outdoors over long duration and distance. Using a pre-built landmark database of the entire exercise area provides precise tracking and eliminates the problem of long term drift inherent in any inertial based navigation platform. Capturing the 3D reconstruction uncertainty of landmark points improves the stability of pose estimation, which is an essential requirement for an augmented reality system.

We showed how to construct this landmark database automatically using sensors mounted on a mobile robot. Finally, we showed how we can take advantage of the cameras mounted on the helmet to estimate the position of the weapon with respect to the helmet.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Department of the Air Force, Eglin AFB and the Office of Naval Research under Contract FA9200-07-D-0045/0016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US Air Force or the Office of Naval Research.

REFERENCES

[Aron'2007] M. Aron, G. Simon, and B. M. O. Use of inertial sensors to support video tracking. Computer Animation and Virtual Worlds, 18, 2007.

[Bleser'2009] G. Bleser and D. Stricker. Advanced tracking through efficient image processing and visual-inertial sensor fusion. Computers and Graphics, 33, 2009.

[Claus'2004] D. Claus and A. Fitzgibbon, Reliable fiducial detection in natural scenes, *ECCV*, pp.469-480, 2004.

[Fontana'2002] Robert J. Fontana and Steven J. Gunderson, Ultra-Wideband Precision Asset Location System, 2002 IEEE Conference on Ultra Wideband Systems and Technologies, Baltimore, MD, May 2002,.

[Foxlin'2003] E. Foxlin and L. Naimark. Vis-tracker:a wearable vision-inertial selftracker. In IEEE Virtual Reality, 2003.

[Hol'2006] J. Hol, T. Schon, F. Gustafsson, and P. Slycke. Sensor fusion for augmented reality. In International conference on information fusion, 2006.

[Jiang'2004] B. Jiang, U. Neumann, and S. You. A robust hybrid tracking system for outdoor augmented reality. In *IEEE Virtual Reality*, 2004.

[Kato'1999] H. Kato and M. Billinghurst, Marker tracking and hmd calibration for a video-based augmented reality conferencing system. *Int'l Workshop on AR*, pp.85-94, 1999.

[Kuipers'1998] J. B. Kuipers. *Quaternions and Rotation Sequences*. Princeton University Press, 1998.

[Mattheis'1987] L. Matthies and S. Shafer. Error modeling in stereo navigation. *IEEE Journal of Robotics and Automation*, 3, 1987.

[Mirzaei'2008] F. M. Mirzaei and S. I. Roumeliotis. A kalman filter-based algorithm for IMU-camera calibration: Observability analysis and performance evaluation. *IEEE Transactions on Robotics*, 24(5), 2008.

[Nister'2007] D. Nister and H. Stewenius, A minimal solution to the generalized 3-point pose problem, *Journal of Mathematical Imaging and Vision*, volume 27, pp.67-79, 2007.

[Nister'2006] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[Oskiper'2007] T. Oskiper, Z. Zhu, S. Samarasekera, and R. Kumar. Visual odometry system using multiple stereo cameras and inertial measurement unit. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[Rehbinder'2003] H. Rehbinder and B. Gosh. Pose estimation using line-based dynamic vision and inertial sensors. *IEEE Transactions on Automatic Control*, 48, 2003.

[Reitmayr'2006] G. Reitmayr and T. Drummond. Going out: robust model-based tracking for outdoor augmented reality. In *International symposium on mixed and augmented reality*, 2006.

[Roumeiotis,2002] S. I. Roumeliotis, A. E. Johnson, and J. F. Montgomery. Augmenting inertial navigation with image-based motion estimation. In *IEEE International Conference on Robotics and Automation*, 2002.

[Roumeliotis'1999] S. I. Roumeliotis, G. S. Sukhatme, and G. Bekey. Circumventing dynamic modeling: Evaluation of the error-state kalman filter applied to mobile robot localization. In *IEEE International Conference on Robotics and Automation*, 1999.

[Saab' 2010] http://saabtraining.com/PDF/PTD_3.pdf

[Schon'2007] T. Schon, R. Karlsson, D. Tornqvist, and F. Gustafsson. A framework for simultaneous localization and mapping utilizing model structure.In *International conference on information fusion*, 2007.

[Se'2006] S. Se, D. Lowe, and J. Jittle. Vision-based global localization and mapping for mobile robots. *IEEE Transactions on Robotics*, 21(3), 2006.

[You'2001] S. You and U. Neumann. Fusion of vision and gyro tracking for robust augmented reality registration. In *IEEE Virtual Reality*, 2001.

[Zhu'2008] Z. Zhu, T. Oskiper, S. Samarasekera, R. Kumar, and H. S. Sawhney. Real-time global localization with a pre-built visual landmark database. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008*