GroundCam: A Tracking Modality for Mobile Mixed Reality

Stephen DiVerdi* Tobias Höllerer†

"Four Eyes" Laboratory
Computer Science Department
University of California, Santa Barbara

ABSTRACT

Anywhere Augmentation pursues the goal of lowering the initial investment of time and money necessary to participate in mixed reality work, bridging the gap between researchers in the field and regular computer users. Our paper contributes to this goal by introducing the GroundCam, a cheap tracking modality with no significant setup necessary. By itself, the GroundCam provides high frequency, high resolution relative position information similar to an inertial navigation system, but with significantly less drift. When coupled with a wide area tracking modality via a complementary kalman filter, the hybrid tracker becomes a powerful base for indoor and outdoor mobile mixed reality work.

Keywords: Anywhere augmentation, vision-based tracking, tracker fusion, mobile mixed reality.

Index Terms: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Motion I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—Motion

1 INTRODUCTION

Traditional mixed reality applications are built on a series of assumptions about the environment they will operate in, often requiring time-consuming offline measurement and calibration for model construction purposes or instrumentation of the environment for tracking. This high start up cost limits the general appeal of mixed reality applications, creating a barrier to entry that discourages potential casual mixed reality users. The goal of *Anywhere Augmentation* is to create a class of mixed reality technologies and applications that require a minimum of setup using cheap, commonly available hardware, bringing the field of mixed reality within the realm of an average computer user.

The choice of tracking technology used in a mixed reality application is heavily dependent on the environment, and its setup and calibration is often one of the time consuming initial steps of application deployment. An overview of the commonly available technologies is presented in Table 1. It is apparent that no single tracking solution exists for the interesting and increasingly common case of wide area, high resolution applications, such as outdoor architectural visualizations. The prevailing solution is to couple a *global* tracker such as GPS, which provides wide area, absolute, low resolution data, with *local* tracking, e.g. from inertial sensors, which provides high resolution, relative and drift prone positioning.

Table 1: A brief comparison of tracking technologies, for typical setups. *range*: size of the region that can be tracked within. *setup*: amount of time for instrumentation and calibration. *resolution*: granularity of a single output position. *time*: duration for which useful tracking data is returned (before it drifts too much). *environ*: where the tracker can be used, indoors or outdoors. All values are expressed accurate to orders of magnitude.

technology	range (m)	setup (hr)	resolution (mm)	time (s)	environ
magnetic [6]	1	1	1	∞	in/out
ultrasound [26]	10	1	10	∞	in
inertial [2]	1	0	1	10	in/out
pedometer [25]	1000	0	100	1000	in/out
optical,					
beacons [23]	10	1	1	∞	in
passive [22]	10	10	10	∞	in
markerless [4]	10	0	10	∞	in/out
hybrid [9]	10	10	1	∞	in
GPS [10]	∞	0	1000	∞	out
beacons [29, 24]	100	10	1000	∞	in/out
WiFi [1]	100	10	1000	∞	in/out
GroundCam	10	0	1	1000	in/out

In this paper, we introduce the GroundCam (consisting of a camera and an orientation tracker - see Figure 4), a local tracking technology for both indoor and outdoor applications. We use the optical flow of a video of the ground to determine velocity, inspired by the workings of an optical mouse. This is related to visual odometry work done in the robotics community [3, 27], but here we apply it to the much less constrained world of human tracking. By itself, the GroundCam provides high resolution relative position information, but is subject to drift due to integration of error over time. From Table 1, it is clear the GroundCam most similarly resembles an inertial tracker, which measures acceleration and integrates twice to get position. The GroundCam is a significant improvement over inertial tracking because its single integration accumulates error much more slowly, maintaining similar small-scale accuracy for a longer period of time.

To address the GroundCam's long term drift, we use a

*e-mail:sdiverdi@cs.ucsb.edu
†e-mail:holl@cs.ucsb.edu

IEEE Virtual Reality Conference 2007
March 10 - 14, Charlotte, North Carolina, USA
1-4244-0906-3/07/$20.00 ©2007 IEEE

complementary Kalman filter to combine the GroundCam with a wide area sensor such as a GPS receiver (see Figure 4), providing better accuracy over large environments. For wide area indoor operation, we simulate the signal from a beacon-based tracker such as the Cricket [24] or Locust Swarm [29], to demonstrate the hybrid performance. These wide area trackers provide periodic stable corrections to compensate for the GroundCam's drift while maintaining its fast and high resolution data.

The advantages of the GroundCam include its favorable performance compared to other local tracking technologies, as well as its general applicability to a variety of mixed reality applications, including outdoor mobile augmented reality and indoor virtual reality. Hybrid indoor / outdoor applications can also use the GroundCam, as it handles large changes in illumination gracefully. Finally, the low cost and ease of construction of the GroundCam make it suitable towards our goal of Anywhere Augmentation, by reducing the barriers to entry for mixed reality applications.

2 Related Work

Related work falls into the following main categories: optical flow-based tracking techniques and hybrid tracking approaches with a focus on pedestrian navigation.

2.1 Optical Flow-Based Tracking

Using optical flow for camera tracking has been explored in many applications. The widest commercial distribution was reached by the modern optical mouse, which uses an LED to illuminate the mouse pad surface for a small camera that tracks texture motion across the visual field, generating a translation vector. For optical mice, the problem is drastically simplified by the assumption that the entire visual field will exhibit a single coherent translation. A similar concept is implemented as part of Haro et al's mobile UI work [13], which uses optical flow from a cell phone camera as a 2D input to mobile application user interfaces. Many concessions must be made due to the phone's limited processing power – most importantly, the motion estimation is limited to one of four cardinal directions and very approximate measures of motion magnitude are used. Given our interest in accurate and robust tracking, we cannot make similar simplifying assumptions for the GroundCam.

Much of the previous work in camera tracking via optical flow is in the field of visual odometry for mobile robotics. A straightforward approach is taken by Lee and Song [18], mounting an optical mouse near the ground on a wheeled robot. While the mouse does provide high quality optical flow information, the fact that it needs to be within a few millimeters of the tracked surface inherently restricts the robot to very smooth terrain. A more sophisticated solution is to mount a camera horizontally on the robot to provide an eye-like view of the world, as in Campbell et al's visual odometry evaluation [3]. While their work tests the performance of visual odometry in terrain that is difficult for robots, such as ice and grass, the ground is still required to be flat and free of distracting influences. The GroundCam's design allows it to be used in complex terrain including obstacles and debris, significant changes in height, and other moving agents.

Se et al's robot [27] handles complex environments by using SIFT features and a three camera stereo system. SIFT features are matched across images from the three cameras to build a 3D map, against which features in subsequent frames are matched. While the results are impressive, the algorithm is also very demanding computationally, operating at 2Hz and restricting their robot to a speed of 0.4m/s. Nistér et al [21] use stereo imagery for visual odometry for ground vehicles, with good accuracy and ability to handle distractions. However, the updates are limited to 10Hz, and necessary temporal filtering takes advantage of the low frequency of accelerations for a ground vehicle. Human tracking requires low latency, high frequency updates for interactive applications.

2.2 High Quality, Wide Area Tracking

None of the methods discussed so far are sufficiently robust and/or precise to work for arbitrary wide area mixed reality applications. Therefore, tracking approaches for such environments are typically of a hybrid nature.

Foxlin and Naimark [9] propose the coupling of inertial sensors with vision-based tracking of fiducial markers that have to be attached to the ceiling or walls around the tracking area. By tracking natural features the GroundCam does not require instrumentation of the environment.

A common approach for tracking pedestrians in the outdoors is to couple GPS tracking with inertial-based dead reckoning to improve update rates and to bridge areas where GPS is unreliable [5]. Relying on inertial sensors as a direct position tracking modality is widely deemed to be of limited use, however, because of the rapid propagation of drift errors due to double integration [2]. Instead, many systems employ inertial sensors as a pedometer, detecting the event of the user taking a step [25, 12, 5]. For indoor navigation, Lee and Mase couple step-detection via inertial sensors with infrared beacons for absolute measurements [17]. For all these hybrid tracking techniques, the GroundCam provides an additional dead-reckoning sensor that could improve accuracy and reliability of the position tracking.

As hybrid tracking systems are often used to address the limitations of individual tracking modalities, there has been extensive research into techniques for optimally coupling these sensors. Foxlin [8] originally used a complementary separate-bias Kalman filter to combine gyroscopes, inclinometers and a compass, while You and Neumann [31] use an extended Kalman filter with separate correction steps for vision and gyroscope updates. Jiang et al [16] combine vision and gyroscope sensors in a more heuristic manner – the gyroscope measurement is used as an initial estimate to limit the vision feature search, and the vision measurement is used to limit the gyroscope drift. Finally, while coupling between

```
loop:
  getFrame()
  undistortFrame()
  if( num_features < max_features )
    findNewFeatures()
  findOpticalFlow()
  findInliers()
  getOrientation()
  computeMotion()
  reportPosition()
```

Figure 1: Pseudocode for the GroundCam algorithm.

sensors is often loose, there is also work in tightly coupled sensors, such as GPS/inertial hybrids [19]. For our hybrid tracker, we loosely couple the GroundCam and GPS units for modularity and simplicity of design.

3 GroundCam

The inspiration for the GroundCam is a desktop optical mouse. A camera is pointed directly at the ground from just above waist height, and the video of the ground moving in front of the camera is used to determine how the camera is moving in the plane of the ground. The result is a 2D position tracker. Depending on the environment the GroundCam is being used in, it could be more useful if directed at the ceiling – for example, an indoor location with a featureless floor but a textured ceiling. Operation is the same in either case.

3.1 Implementation

The GroundCam takes a few straightforward steps to compute user motion. Pseudocode of this algorithm can be found in Figure 1. Since features are lost and must be added again each frame, there is no explicit initialization step – instead, the first frame is treated as the case where all the features were lost in the previous frame. This means the GroundCam can recover even in cases of total image loss, such as sudden extreme dark or bright conditions, without any user intervention. For the algorithms used in the GroundCam, standard implementations from the OpenCV image processing library [15] are used, unless stated otherwise.

3.1.1 Undistortion

Offline intrinsic camera calibration is done using Zhang's procedure [32]. The distortion coefficients from this process are used to correct the resulting artifacts in the video frames by creating a corresponding undistortion image warp that is applied to each frame – this allows us to use image distances as direct measurements of distances in the scene. However, for cameras with a narrow field of view, the distortion effect is small enough that it does not produce a significant effect and undistortion is unnecessary, saving CPU cycles – for example, we do not undistort the video for the camera from Figure 4, which has a field of view of 12.2 degrees.

3.1.2 Feature Detection

In our system, features are small regions of image texture. Good features for tracking are selected from the video frames by Shi and Tomasi's algorithm [28], which finds a set of all features of a certain quality and then greedily selects features from the set that are not within a minimum distance of the already selected features. After the initial set of features are found, new features are introduced with the same technique as features are lost.

3.1.3 Feature Tracking

Features are tracked frame to frame using the image pyramid based optical flow algorithm of Lucas and Kanade [20]. A hierarchy of images at different resolutions are used to efficiently match texture features from one frame with the most similar region in another frame. If the similarity between these two regions is below a threshold, the feature is considered lost and is removed from the set. This can happen when a feature goes outside the field of view, or when changes in illumination or occlusion occur. Each feature is tracked independently of one another, so their motion may not be (and in most cases is not) uniform. This is a strength of the technique, as distractors can be accounted for so long as overall they represent a minority of the viewable scene.

3.1.4 Coherent Motion Estimation

Coherent motion must be extracted from the set of features successfully found in consecutive frames, discarding the influence of outliers. We implemented the RANSAC algorithm [7] to accomplish this task. Only one sample is necessary to estimate the image's 2D translation. Other samples are tested against this estimate by separately thresholding the differences in magnitude and orientation. Once the final set of inliers is found, the image motion estimate is computed by taking the average of all the good samples. In the event that a consensus is not reached, a fallback estimate is computed as the average of all the samples.

The computation to get world motion in real units from the image motion in pixels is straightforward. The camera is assumed to be perpendicular to the ground at some uniform height (measured offline). For a known height in meters H, camera horizontal field of view F, and camera width in pixels P, the conversion factor from pixels is

$$\frac{2H}{P}\tan\left(\frac{F}{2}\right) \tag{1}$$

A 640x480 image from our camera with a field of view of 12.2 degrees, mounted at 1.1m (just above waist height), yields a factor of 0.37mm per pixel.

Our implicit assumption that the conversion between image distance and physical distance can be represented by a single scale factor is not actually correct. For different regions of the image, the distance from the camera to the ground varies, even assuming a flat ground and perfectly orthogonal viewing direction. The scale factor we computed

is therefore not correct outside of the center of the field of view. For our camera with a 12.2 degree field of view, a simple calculation shows that a 0.5% error is introduced between computations at the center and the perimeter of the image. This is small enough that we can safely ignore it for our purposes.

3.1.5 World Coordinate Transformation

Our motion estimate is computed in the camera's frame of reference. In order to convert it to the world's coordinate system, we need to know the absolute orientation of the camera. An InterSense InertiaCube2 orientation tracker is used to obtain this information. A quick offline calibration is done to orient the InertiaCube2's output by obtaining angles for north, east, south and west. During operation, the detected angle is linearly interpolated between these computed values to get the world stabilized camera orientation. The motion vector is then transformed by this orientation to yield the final, world stabilized motion estimate.

3.2 Discussion

Numerous experiments using the GroundCam have yielded the following insights into its setup and operation in real world conditions.

3.2.1 Performance

The output of the GroundCam is essentially a linear velocity measurement – some distance traveled over a small unit time. Therefore, integration is necessary to use it as a position tracker. However, it compares favorably to the primary alternative, a linear accelerometer, as the acceleration data requires double integration to yield position, and so accumulates error much faster. Single integration means our drift over time is drastically reduced. Also similar is the use of a pedometer to track walking motion, which uses a known stride length and counts steps to arrive at a position estimate. However, pedometers are limited to the resolution of a stride, and can drift significantly when the user walks with steps of unusual stride (e.g. due to terrain considerations like stairs, or from repeated small steps when carefully adjusting ones position).

The limiting factors in the GroundCam's feature tracking are the camera's image quality and framerate and the size of the visible ground region. Good lighting and good optics improve the performance of the optical flow algorithm significantly – optics improve image quality and bright light lowers the necessary exposure time, reducing image noise and motion blur (daylight or bright office illumination is generally sufficient). For our setup, we use a Unibrain Fire-i 400 camera with a 12.2 degree field of view lens mounted at 1.1 meters, which yields a ground section of 0.24m by 0.18m. For half the features to still be visible, the ground can move at most half of this region, 0.09m along the y-axis, in the time of a single frame. At 10fps, that is equivalent to a speed of 0.88m/s, at 15fps, 1.18m/s, and at 20fps, 1.76m/s. Since forward motion is most common, mounting the camera rotated 90 degrees (portrait vs. landscape), gives 0.24m of visible ground along the walking dimension, and results in a trackable speed of 1.32m/s at 10fps, 1.76m/s at 15fps, or 2.35m/s at 20fps. Average walking speed is 3mph or 1.34m/s, and we consistently get between 15fps and 20fps, so this is sufficient for basic walking behavior. Fast walking or running cause sufficient jitter of the camera's orientation, resulting in significant motion blur and noisy apparent motion, such that they cannot be accurately tracked in any case.

3.2.2 Feature Selection

Our choice of 50 tracked features bears justification. A manual comparison was done of the GroundCam's coherent motion estimate output for different target numbers of tracked features, from 25 to 200 (since some number of features are lost each frame, the actual set of features present in two consecutive frames is less than the target) . At 25, there were few enough points that a coherent estimate often was not possible, or else it was very likely to get distracted by random noise over low texture terrain. At 100 and above, the probability of achieving a coherent estimate was very high, but it was necessary to increase the number of inliers required for RANSAC to succeed, so the overall gain in detection was small. However, the additional CPU drain in tracking, executing RANSAC, and replenishing lost features was significant. 50 features is a compromise between CPU cost and likelihood of detecting coherent motion. It may be possible to fine tune this parameter for particular known types of terrain, but we preferred a single static value.

3.2.3 Orientation Estimation

The need for an orientation tracker is not necessarily clear in light of research such as Davison's single camera SLAM [4]. In his work, the camera's 6DOF pose is completely determinable from the video stream. However, the dependable high contrast of the texture in his work makes the tracking much more reliable than for the GroundCam. Over high contrast terrain such as well-lit grass or gravel, there may be enough information to extract the camera's orientation as well, but on terrain such as concrete, asphalt, or hard dirt there is enough optical flow noise that it is difficult to reliably extract the translation motion estimate. Techniques to improve the quality of feature tracking for these types of terrain could make the individual feature motion information reliable enough for full 6DOF pose estimation.

3.2.4 Tracking Distractions

There are a number of possible distractions that can reduce the accuracy of the tracking result. The user's lower legs and feet may appear in the camera's field of view, which can create strong enough optical flow to influence the motion estimate. It would be possible to create a color model of the lower legs and feet, which is likely different from the ground terrain, and use it to mask them out of the ground image before tracking features on it. Alternately, proper mounting of the camera (e.g. on the back of a backpack containing the

wearable computer) with a narrow field of view alleviates this problem by keeping the feet out of the video.

Motion of the user's shadow can have a similar effect, if the leg shadows are moving across the camera's view, by creating many strong features on the shadow boundary that move separately from the ground. Mounting the camera on the the front or back diminishes the effect, as the forward-backward motion of legs creates shadows with much less motion in those cases. It would also be possible to use image processing to remove large scale illumination changes while keeping small, local texture, at additional CPU cost.

Changing illumination when moving from a well-lit area into a poorly-lit one, such as crossing into the shadow of a building can create temporary confusion as the camera's exposure setting automatically adjusts, depending on the contrast and sharpness of the shadow. This is only a serious problem when the contrast creates under or over saturation of regions of the image, which then do not have trackable texture. As soon as the camera's exposure adjustment compensates (up to one second), tracking resumes as normal.

Finally, changes in height of the ground plane, either from structures like stairs, or random debris on the ground (e.g. rocks) may introduce error in the conversion between motion in pixels to meters. Potentially, the use of SIFT features would allow determination of changing heights, such as when going up or down stairs, which could then be used to improve the coherent motion estimate. For objects or debris of static height above the ground plane, more reliable feature tracking would be necessary to identify coherent patches of motion of different velocity than the majority of the image and then extract a height estimate.

However, all of these possible distractions combined do not, in general, significantly impact the quality of the tracking result. They are all short, temporary effects that introduce small amounts of random noise. The result is that they cause the integrated position to drift slightly faster than it would under ideal conditions, but this has not shown to be a problem.

3.2.5 Camera Parameters

The camera's height, angle, and field of view, are all important considerations. Since the GroundCam aims to track the position of the user while maintaining its position relative to the ground, it must be mounted on the hips or torso. It should not restrain the user from moving their arms or doing their work, so the back is best. For wearable systems that include a backpack-like computer, the back of such a device is an ideal location (see Figure 4).

The height of the camera is important in conjunction with the field of view – the resulting size of the viewable ground region affects the maximum speed that can be tracked, as discussed earlier. It also affects the size of texture features that can be used for tracking. Keeping the viewable ground region small has the advantage of reducing the potential for distractors such as feet to interfere with the tracking. On the other hand, it increases the potential for distractors to take over the majority of the field of view and significantly confound the tracking. These tradeoffs must be considered per-application.

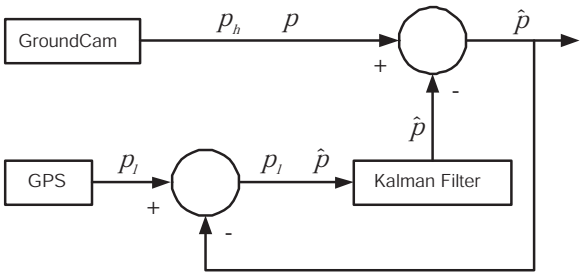

Figure 2: System diagram of the complementary Kalman filter. A Kalman filter is used to update an error between the current GroundCam estimate and the GPS absolute position. While the Kalman filter is updated infrequently (1Hz), a new position estimate is generated for each GroundCam update (30Hz).

We chose to point the camera straight down, perpendicular to the ground. This choice has a few nice properties – first, it makes the motion estimation easy to compute and the matching of samples to an estimate in the RANSAC algorithm is similarly easy. Second, pointing straight down minimizes the total volume of the viewing frustum of the camera, which means there is less volume for distractors to intrude. Third, this orientation makes tracking easier as features exhibit the smallest change in appearance moving across the field of view. Finally, since the camera is not rigidly held with respect to the ground, its orientation is likely to change slightly during operation – small changes from this orientation will have a smaller impact on the assumptions made than from other orientations. For example, if the camera were to become misaligned by 5 degrees, a simple calculation shows the error in the motion estimate would be $\leq 2\%$, which is small enough that it can be ignored.

4 Hybrid Tracking

The GroundCam by itself is not a sufficient wide area tracking solution because it tends to drift over prolonged operation. Instead, it is most appropriately used in concert with a wide area tracker like a GPS receiver. This loose coupling is achieved with a complementary Kalman filter.

4.1 Complementary Kalman Filter

Our complementary Kalman filter design is inspired by Foxlin's work on orientation tracker filtering [8]. The underlying concept is to filter the error signal between two sensors, rather than filtering the actual position estimate (see Figure 2).

The signal from the GroundCam is high frequency (30Hz), high resolution (1mm), and includes small random errors (10mm) and large systematic errors (drift is unbounded over time). There are two main sources of error – random errors in the motion estimates per update, and random underestimation of motion when RANSAC fails to find a coherent estimate. These errors accumulate over time due to integration

of the GroundCam signal. The signal from a standard GPS receiver is low frequency (1Hz), medium resolution (10cm), and includes medium random and systematic errors (5m). The main source of the error is due to changing atmospheric conditions which delay the signals from GPS satellites differently, creating apparent differences in position. Generally, this error is randomly distributed around the true position, but prevailing weather conditions such as cloud cover, or view obstructions such as buildings, can create systematic errors in the signal over extended periods of time.

Ideally, the filtered output will be available at the high frequency and high resolution, which the complementary Kalman filter achieves with minimal processor load. We can model the error between the two signals as a smoothly varying random process with a Kalman filter, and then use the filtered error signal to correct the GroundCam signal on the fly.

Let p_h be the high frequency signal from the GroundCam, and p_l the low frequency signal from a GPS receiver. p is the ground truth position, $\hat{p}$ is the estimated position, $\delta p = p - p_h$ and $\delta \hat{p}$ is the estimated error signal. Since our filter operates on 2D data, p is actually the vector $[x,y]^T$. Within the Kalman filter, there are 6 process dimensions and 2 measurement dimensions. Filter variable names are standard as used in [30].

$$\mathbf{x} = \begin{bmatrix} \delta p \\ \delta \dot{p} \\ \delta \ddot{p} \end{bmatrix} \tag{2}$$

$$\mathbf{z} = \begin{bmatrix} \delta p \end{bmatrix} \tag{3}$$

$$A = \begin{bmatrix} 1 & \Delta t & \frac{1}{2}\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

$$H = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \tag{5}$$

B and **u** are both not used, and thus zero. Q and R are empirically determined depending on the particular sensor being coupled with the GroundCam, and P is initially set so measurements are preferred at startup.

The result of a complementary filter setup such as this is that for each new high frequency update, only a prediction and then subtraction is necessary, making the processor load very low for the frequent step. The expensive correction step is computed once per low frequency update.

4.2 Potential for Coupling

There are a number of possible wide area trackers that could be integrated with the GroundCam in this manner, depending on the needs of the particular system. GPS is a straightforward choice for outdoor applications, as its signal is commonly available and sensors are cheap. Applications without a clear view of the sky however, such as dense urban environments or indoors, must consider alternative solutions. In these cases, a cheap and easily deployable beacon-based system, e.g. on RF, ultrasound, or infrared basis [29, 24, 11, 14], may be more appropriate. Such systems provide position information in the sense that they identify which discrete region the user currently occupies. This information would be sufficient for applications such as audio annotations or situated content, but for visual overlays or immersive virtual content, coupling with a more accurate tracker like the GroundCam is necessary.

The coupling of the GroundCam with another sensor may be done differently as well, for different needs. For instance, a common problem with GPS signals is that while the user is standing still, error in the GPS signal will make it appear as though the user is moving slowly. This drift can make operations that require stationary actions very difficult. The GroundCam, on the other hand, is very good at determining when the user is standing still and could be used as a binary walking / standing behavior classification, to selectively ignore GPS updates.

5 Results

Figure 5 shows a typical run using the GroundCam and GPS hybrid tracking system for approximately 90 seconds. The path includes avoiding obstacles and going up and down steps, with wood, gravel and concrete terrain. As expected, the GroundCam exhibits some drift, partially from random errors in the motion estimate but also from updates where a coherent estimate cannot be generated. These errors cause different effects in the GroundCam path – random errors make the path less smooth, while missing coherent estimates create a shortening effect. However, the coupling with the GPS signal eliminates the effect of the GroundCam drift. Of particular importance is the much smoother quality of the filtered signal than the raw GPS signal, which makes the hybrid tracker very appropriate for mixed reality applications.

For comparison purposes, the run in Figure 6 includes a hand-labeled ground truth – a rectangular path of approximately 18m x 12m over 81 seconds on a residential street. The terrain is concrete and asphalt, which have lower contrast textures and are more prone to noise in the error estimates. For this particular trial, our GPS receiver experienced very little random noise, but did exhibit a significant drift overall, due to our GPS unit not receiving a WAAS signal at our location. While our filtered path stays close to the GPS signal, we cannot correct for systematic errors in the GPS position, which are propagated into our tracking result. In most US locations, the presence of a WAAS signal will improve the quality of the GPS data and subsequently improve the filtered data as well.

5.1 Slip Compensation

The problem of RANSAC not reaching a consensus is analogous to the problem of slipping wheels in odometry of

wheeled vehicles, and results in estimated paths that are much shorter than ground truth. Certain types of terrain are more prone to this sort of error (see Figure 3). Low-contrast terrains like concrete were much more prone to slipping than high-contrast terrains such as grass.

We made a simple attempt to compensate for some of this error, which we call *slip compensation*. The error is proportional to the rate at which RANSAC does not produce a coherent estimate, or the *slip rate*. Based on the slip rate over a short window of time, a successful coherent estimate is scaled to compensate for the missed estimates (e.g. if the slip rate is $s = 0.8$, then a coherent estimate is scaled by $(1-s)^{-1} = 5.0$). Figure 7 clearly shows that slip compensation helps achieve the appropriate scale of the GroundCam signal.

5.2 Beacon-based Wide Area Sensors

To demonstrate the usefulness of the GroundCam in concert with wide area sensors other than GPS, we simulated a discrete beacon-based wide area sensor signal (similar in concept to the Cricket [24] and Locust Swarm [29] projects). We used ground truth to trigger a periodic signal that identified which discrete region the user currently occupied, on a rectilinear grid of 6m cells (roughly room-size). This coarse wide area signal was used in place of the GPS signal in the complementary Kalman filter.

Figure 8 shows that the beacon-based signal provides a measure of drift-correction that improves the GroundCam's raw result. For a longer path, the GroundCam drift would result in significant divergence from ground truth, while the beacon-based signal would make sure the filtered output stays within certain bounds of the true position.

6 Conclusion

We have presented the GroundCam tracking modality, a vision-based local tracker with high resolution, good short-term accuracy, and an update rate appropriate for interactive graphics applications. We have also demonstrated the feasibility of a hybrid tracker, coupling the GroundCam with a GPS receiver, as well as a discrete beacon-based wide area sensor. In our trials, the GroundCam compares favorably to other similar tracking modalities, and we are currently integrating it into our mobile mixed reality platform for experience in actual application scenarios. Towards the goal of Anywhere Augmentation, the GroundCam is cheap, readily available, and requires almost no time to setup in a new environment, for high-quality tetherless tracking in mixed reality applications.

Acknowledgements

This research was in part funded by a grant from NSF IGERT in Interactive Digital Multimedia #DGE-0221713 and a research contract with the Korea Institute of Science and Technology (KIST) through the Tangible Space Initiative Project.

References

[1] P. Bahl and V. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. 2:775–784, 2000.

[2] B. Barshan and H. Durrant-Whyte. Inertial navigation systems for mobile robots. *Transactions on Robotics and Automation*, 11(3):328–342, 1995.

[3] J. Campbell, R. Sukthankar, and I. Nourbakhsh. Techniques for evaluating optical flow for visual odometry in extreme terrain. In *Proceedings of the International Conference on Intelligent Robots and Systems*, volume 4, pages 3704–3711, 2004.

[4] A. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proc. IEEE ICCV '03*, Oct. 2003.

[5] L. Fang, P. Antsaklis, L. Montestruque, M. McMickell, M. Lemmon, Y. Sun, H. Fang, I. Koutroulis, M. Haenggi, M. Xie, and X. Xie. Design of a wireless assisted pedestrian dead reckoning system - the NavMote experience. *Transactions on Instrumentation and Measurement*, 54(6):2342–2358, 2005.

[6] FastTrack, September 2006. Polhemus, http://www.polhemus.com/.

[7] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395, 1981.

[8] E. Foxlin. Inertial head-tracker sensor fusion by a complementary separate-bias kalman filter. In *Proceedings of the Virtual Reality Annual International Symposium*, pages 184–194, 1996.

[9] E. Foxlin and L. Naimark. VIS-tracker: a wearable vision-inertial self-tracker. In *Proceedings of Virtual Reality*, pages 199–206, 2003.

[10] I. Getting. The global positioning system. *IEEE Spectrum*, 30(12):36–47, Dec. 1993.

[11] D. Hallaway, T. Höllerer, and S. Feiner. Coarse, inexpensive, infrared tracking for wearable computing. In *Proceedings of the International Symposium on Wearable Computers*, pages 69–78, 2003.

[12] D. Hallaway, T. Höllerer, and S. Feiner. Bridging the gaps: Hybrid tracking for adaptive mobile augmented reality. *Applied Artificial Intelligence Journal, Special Issue on AI in Mobile Systems*, 18(6):477–500, 2004.

[13] A. Haro, K. Mori, T. Capin, and S. Wilkinson. Mobile camera-based user interaction. In *Proceedings of the International Conference on Computer Vision Workshop on Human Computer Interaction*, pages 79–89, 2005.

[14] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster. The anatomy of a context-aware application. In *Proceedings of the International Conference on Mobile Computing and Networking*, pages 59–68, 1999.

[15] Intel Corporation. *Open Source Computer Vision Library Reference Manual*. December 2000.

[16] B. Jiang, U. Neumann, and S. You. A robust hybrid tracking system for outdoor augmented reality. In *Proceedings of Virtual Reality*, pages 3–10, 2004.

[17] S. Lee and K. Mase. A personal indoor navigation system using wearable sensors. In *Proceedings of the International Symposium on Mixed Reality*, pages 147–148, 2001.

[18] S. Lee and J. Song. Mobile robot localization using optical flow sensors. *International Journal of Control, Automation, and Systems*, 2(4):485–493, 2004.

[19] Y. Li and J. Wang. Low-cost tightly coupled GPS/INS integration based on a nonlinear kalman filtering design. In *Institute of Navigation National Technical Meeting*, 2006.

[20] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[21] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23(1), 2006.

[22] I. Poupyrev, D. Tan, M. Billinghurst, H. Kato, H. Regenbrecht, and N. Tetsutani. Developing a generic augmented-reality interface. *Computer*, 35(3):44–50, March 2002.

[23] Precision position tracker, 2006. WorldViz, http://www.worldviz.com/.

[24] N. Priyantha, A. Chakraborty, and H. Balakrishnan. The cricket location-support system. In *Proceedings of the International Conference on Mobile Computing and Networking*, pages 32–43, 2000.

[25] C. Randell, C. Djiallis, and H. Muller. Personal position measurement using dead reckoning. In *Proceedings of the International Symposium on Wearable Computers*, pages 166–173, 2003.

[26] C. Randell and H. Muller. Low cost indoor positioning system. In *Proceedings of Ubiquitous Computing*, pages 42–48, 2001.

[27] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the International Conference on Robotics and Automation*, pages 2051–2058, 2001.

[28] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1994.

[29] T. Starner, D. Kirsch, and S. Assefa. The locust swarm: An environmentally-powered, networkless location and messaging system. In *Proceedings of the International Symposium on Wearable Computers*, pages 169–170, 1997.

[30] G. Welch and G. Bishop. An introduction to the kalman filter. *SIGGRAPH Course Notes*, 2001. Course 8.

[31] S. You and U. Neumann. Fusion of vision and gyro tracking for robust augmented reality registration. In *Proceedings of Virtual Reality*, pages 71–78, 2001.

[32] Z. Zhang. A flexible new technique for camera calibration. *Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

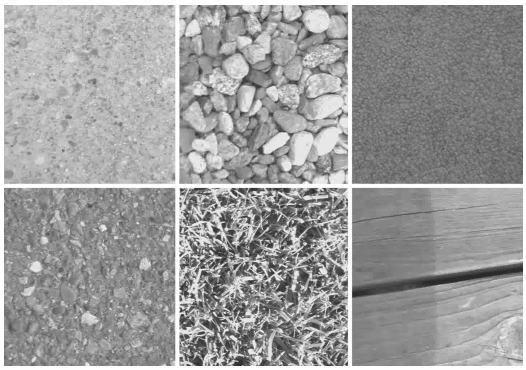

Figure 3: Different types of terrain with example slip rates. *left to right, top to bottom*: concrete (80%), gravel (32%), carpet (48%), asphalt (65%), grass (20%), wood (24%). Slip rates depend on speed, jitter, lighting and debris, in addition to texture contrast.

Figure 4: A wearable computer setup for GroundCam tracking. We use a Unibrain Fire-i 400 camera, an InterSense InertiaCube2 orientation tracker, and a Garmin GPS 18 receiver. Inside the backpack is a Dell Precision M50 laptop.

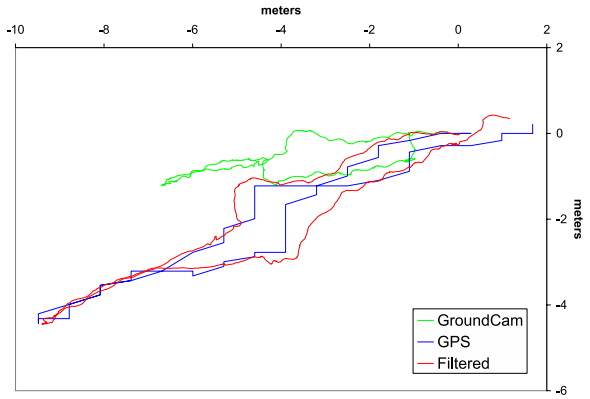

Figure 5: A trial run of the GroundCam coupled with GPS. The run was 90 seconds in duration, over wood, gravel and concrete terrain, and included avoiding obstacles and going up and down stairs. The slip rate was 30%.

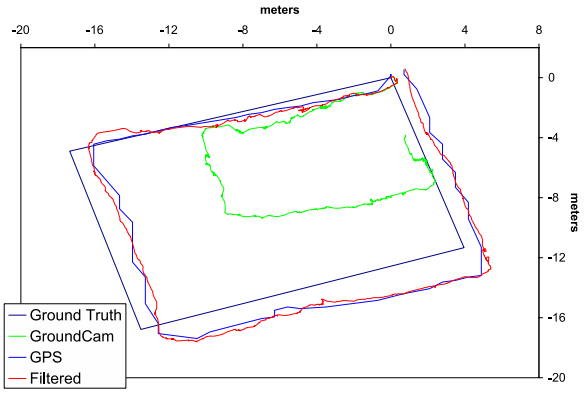

Figure 6: A trial run of the GroundCam coupled with GPS, with hand-labeled ground truth. The run was 81 seconds long, over concrete and asphalt, along a rectangle 18m long by 12m wide. The slip rate was 80% and RMS errors for the GroundCam, GPS, and filtered signals are 5.5m, 1.9m, and 1.9m respectively.

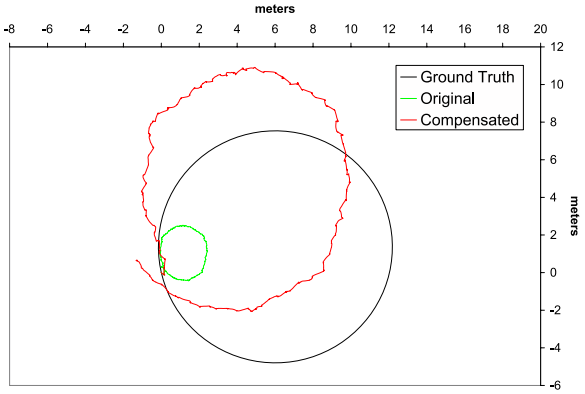

Figure 7: A trial run of the GroundCam with and without slip compensation, with hand-labeled ground truth. The trial was 72 seconds in duration over asphalt, and had a slip rate of 63%. Originally, the RMS error was 7.0m; with slip compensation, the RMS error is 4.8m.

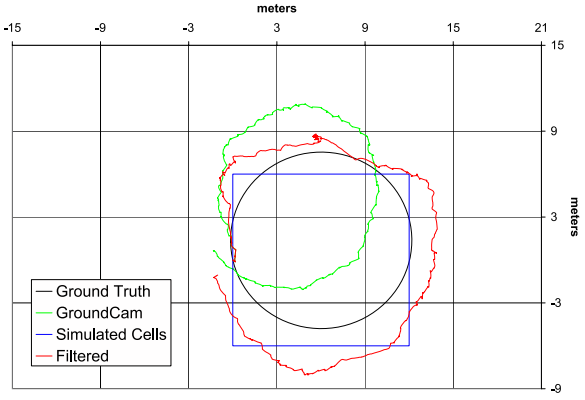

Figure 8: A trial run of the GroundCam (with slip compensation) with a simulated beacon-based wide area sensor in place of the GPS signal. The RMS errors of the GroundCam, beacon signal, and filtered signal are 4.6m, 2.3m, and 1.9m respectively.