Learning Activity-Based Ground Models from a Moving Helicopter Platform*

Andrew Lookingbill, David Lieb, David Stavens, and Sebastian Thrun
Stanford AI Lab
Stanford University
Stanford, CA 94305
{apml,dflieb,dstavens,thrun}@stanford.edu

***Abstract*— We present a method for learning activity-based ground models based on a multiple particle filter approach to motion tracking in video acquired from a moving aerial platform. Such models offer a number of potential benefits. In this paper we demonstrate the ability of activity-based models to improve the performance of an object motion tracker as well as their applicability to global registration of video sequences.**

***Index Terms*— Computer Vision, Machine Learning, Object Tracking, Particle Filters, Activity Maps.**

I. Introduction

In recent decades, the problem of acquiring accurate ground models, or maps, has become the focus of a number of different research communities. Photogrammetry investigates the acquisition of models from remote imaging sensors flown on high-aerial aircraft or satellites [1]. Many roboticists concern themselves with the acquisition of maps from the ground, using mobile robots operated indoors [2], outdoors [3], underwater[4], or in the subterranean world [5].

The vast majority of techniques, however, address the acquisition of *static* models. Moving entities, such as cars, bicyclists, and pedestrians, are usually considered irrelevant to the mapping problem. The thrust of our research is the acquisition of activity-based models, which are models that characterize places based on the type of motion activities that occur. For example, the activities found on roads differ from those found on sidewalks, and even among roads motion characteristics vary significantly. Accurate activity-based ground models offer a number of potential benefits: they can help us understand traffic flow; they can assist unmanned ground vehicles in navigating autonomously (e.g., guide them to stay off a busy road), and they can help us spot activity-related change and abnormalities. Good activity models also facilitate the tracking of individual moving objects, as we shall show in this paper.

The acquisition of activity-related models has been addressed previously. For example, Makris and Ellis use video from surveillance cameras to develop an activity-based model of entry points, exit points, paths, and junctions within a scene [6]. However, their approach assumes a static sensor platform—which greatly facilitates the detection and tracking of moving entities. Stauffer and Grimson also use a static sensor forest to track motion, learn patterns of activity at a site, and classify the observed activities [7]. This approach allows them to identify abnormal behaviors in the scene.

Fig. 1. The Stanford Helicopter is based on a Bergen Industrial Twin platform and is outfitted with instrumentation for autonomous flight (IMU, GPS, magnetometer, PC104). In the experiments reported here we replaced the onboard laser with a color camera.

The problem addressed here is the acquisition of activity-based ground models from a moving platform, such as a helicopter. Our system has been used with the Stanford helicopter shown in Fig. 1. This approach transforms video acquired by the helicopter, and other moving platforms, into probability distributions that characterize the frequency, speeds, and directions of moving objects on the ground, for each x-y location on the ground. To obtain such activity maps, our approach uses a pipeline of techniques for reliably extracting tracks and updating the map statistics. Our algorithm performs feature tracking in the image plane, followed by an optical flow analysis that uses EM to identify features that are likely moving on the ground, similar to the approach taken by Jung and Sukhatme [9]. We then apply multiple particle filters which are spawned, merged, and killed in a manner akin to that proposed by Vermaak, Doucet, and Pérez to reliably identify multiple moving objects on the ground [10]. The resulting tracks from the particle filters are fed into a histogram that characterizes the probability distribution over speeds and orientations of motions on the ground. This probability histogram constitutes the learned activity map. To illustrate the utility of the activity map, we leverage it into an improved particle filter tracker and apply it to the problem of global image registration.

*The authors gratefully acknowledge financial support through the DARPA MARS Program (contracts N66001-01-C-6018 and NBCH1020014).

II. Learning Activity Maps from a Moving Platform

A. Feature Tracking

The first step of our approach involves identifying appropriate features in the camera image and tracking them over multiple frames. In the work of Burt et al., an early approach to this problem involved mimicking the foveation and tracking of a human eye [8]. In our approach features are first identified using an algorithm by Shi and Tomasi [11], which selects unambiguous feature points by finding regions in the image containing large spatial image gradients in two orthogonal directions. A sample of features found by this algorithm, in an image acquired by our helicopter, is shown in Fig. 2a.

The tracking of features is then achieved using a pyramidal implementation of the Lucas-Kanade tracker [12]. This approach forms image pyramids consisting of filtered and subsampled versions of the original images. The displacement vectors between the feature locations in the two images are found by iteratively maximizing a correlation measure over a small window, from the coarsest level up to the original level. The result of tracking features is shown in Fig. 2b. The optical flow of a number of features, tracked through consecutive images and indicated by small arrows in the direction of the flow, is shown.

B. Identifying Moving Objects on the Ground

The principal difficulty of interpreting the optical flow arises from the fact that most of the flow is caused by the platform's ego-motion. The flow shown in Fig. 2b is largely due to the helicopter's own motion; the only exception is the flow associated with the dark vehicle in the scene.

Our approach uses the EM algorithm to identify the nature of the flow. Let $\{x_i, y_i, x'_i, y'_i\}$ be the set of features returned by Lucas-Kanade, where (x_i, y_i) corresponds to image coordinates of a feature in one frame, and (x'_i, y'_i) corresponds to the image coordinates of that feature in the next frame. The displacement between these two sets of coordinates is the velocity of a feature relative to the camera plane (but not the ground!). The probability that $\{x_i, y_i, x'_i, y'_i\}$ corresponds to a moving object on the ground is now calculated using the EM algorithm. Specifically, we define the binary variable c_i that indicates whether the i-th feature is moving. Initially, we set $c_i = 0$ for all i, meaning that all features are assumed to be non-moving. The flow represented by $\{x_i, y_i, x'_i, y'_i\}$ is then used to estimate the image plane transformation that results from ego-motion of the platform. We represent the image plane transformation with an affine model that captures translation, rotation, scaling, and shearing. Due to the small amount of camera motion between individual frames, and the small depth of field of the scene relative to the platform altitude, an affine transformation is a reasonable approximation in most cases. For each point (x_i, y_i) the affine transformation determines its position (x'_i, y'_i) in the subsequent frame:

$$\begin{bmatrix} x'_i & y'_i \end{bmatrix} = \begin{bmatrix} 1 & x_i & y_i \end{bmatrix} \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{bmatrix} \quad (1)$$

Using the set of feature correspondences $\{x_i, y_i, x'_i, y'_i\}$, the linear least squares solution provides the optimal affine parameters $\vec{a}$ and $\vec{b}$.

The key to the identification of moving features is now the E-step: Based on the estimated image plane transformation, our approach calculates the expectation of the binary variable c_i:

$$p(c_i = 1 \mid \vec{a}, \vec{b}) = \eta \cdot const \quad (2)$$

$$p(c_i = 0 \mid \vec{a}, \vec{b}) = \eta \cdot \exp\left\{-\frac{1}{2}\vec{D}^T \Sigma_D^{-1} \vec{D}\right\} \quad (3)$$

where $\vec{D} = \left[\begin{pmatrix} x'_i \\ y'_i \end{pmatrix}^T - \begin{pmatrix} 1 & x_i & y_i \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{pmatrix} \right]$ and η is a normalization factor. The matrix Σ_D is a diagonal matrix of size 2-by-2, containing variances for the x and y components. The subsequent M-step iterates the calculation of the model parameters, but now weighted by the expectations calculated in the E-step. A small number of iterations then leads to an improved ego-motion estimate and, more importantly, an estimate of the probability that a feature is moving, $p(c_i)$.

Fig. 2c shows the result of the EM: The flow vectors shown there as small white arrows all correspond with high likelihood to a moving object. In this example, our algorithm correctly identifies the features associated with the vehicle as moving, whereas most features corresponding to static objects have been identified correctly as static (and are therefore omitted in Fig. 2c).

C. Tracking Moving Objects with Particle Filters

Unfortunately, the data returned by the EM analysis is still too noisy for learning activity-based maps. Our affine model assumes an orthographic projection, and is therefore, in general, insufficient to model all possible platform motion. In addition, some features appear to have a high probability of belonging to moving objects due to association error in the Lucas-Kanade algorithm. The resulting activity map would then show high activity in areas where our affine assumption breaks down or Lucas-Kanade errs.

To improve the quality of the tracking, our approach employs multiple particle filters. This approach is capable of tracking a variable number of moving objects, spawning an individual particle filter for each such object. We chose to experiment with particle filters because of the ease of implementation. Let $(s_k^{[m]} \;\; v_k^{[m]})^T$ be the the m-th particle in the k-th particle filter (corresponding to the k-th tracked object). Note: throughout this paper s_i will refer to a feature's coordinates and v_i to its velocity. The prediction

Fig. 2. (a) Features identified using an algorithm by Shi and Tomasi [11]. (b) Optical flow based on a short image sequence, for an image containing a moving object (dark car). (c) The "corrected" flow after compensating for the estimated platform motion, which itself is obtained from the image flow. The reader may notice that this flow is significantly higher for the moving car. These images were acquired with the Stanford helicopter.

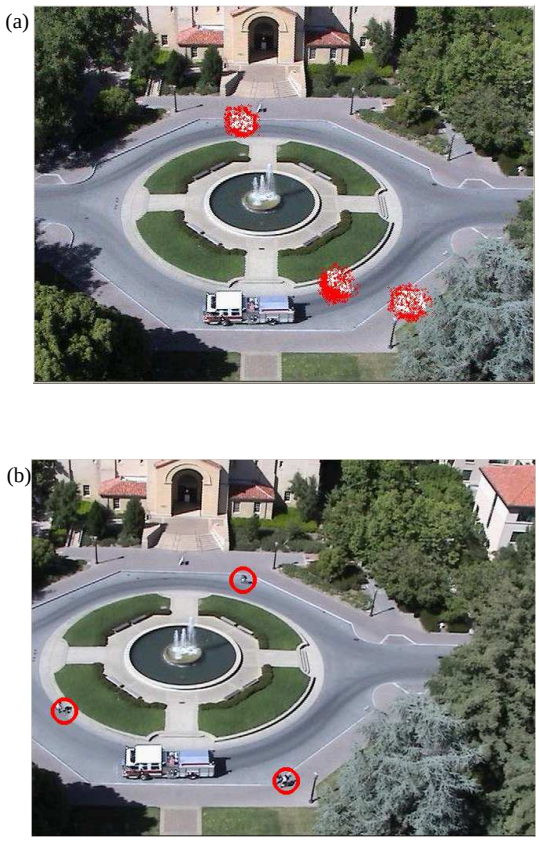


Fig. 3. (a) Multiple particle filters, used for tracking multiple moving objects on the ground. Shown here is an example of tracking three moving objects on the ground, a bicyclist and two pedestrians (the truck in the foreground is not moving). (b) the center of each particle filter in a different frame in the sequence clearly identifies all moving objects.

step for this particle assumes Brownian motion:

$$\begin{pmatrix} s_k^{[m]} \\ v_k^{[m]} \end{pmatrix} \longleftarrow \begin{pmatrix} 1 & \delta t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s_k^{[m]} \\ v_k^{[m]} \end{pmatrix} + \begin{pmatrix} 0 \\ \varepsilon \end{pmatrix} \quad (4)$$

where ε is a random variable modeling the random changes in vehicle velocity, with zero mean and covariance σ. The importance weights are set according to the motion extracted in the previous step. Specifically,

$$w^{[m]} = \sum_i p(c_i) \exp\{\gamma\} \quad (5)$$

where

$$\gamma = -\frac{1}{2}\left[\begin{pmatrix} s_k^{[m]} \\ v_k^{[m]} \end{pmatrix} - \begin{pmatrix} s_i \\ v_i \end{pmatrix}\right]^T \Sigma_w^{-1} \left[\begin{pmatrix} s_k^{[m]} \\ v_k^{[m]} \end{pmatrix} - \begin{pmatrix} s_i \\ v_i \end{pmatrix}\right],$$

$(s_i \quad v_i)^T$ are the motion tracks extracted as described in the previous section, and $p(c_i)$ are the corresponding expectations. The matrix Σ_w is a diagonal matrix of size 4-by-4, with two variances for the noise in location, and two for the noise in velocity. This matrix essentially convolves each track $(s_i \quad v_i)^T$ with a Gaussian with covariance Σ_w.

New particle filters are started if at the border of the camera field a large number of features with high probability $p(c_i)$ exist that are not associated with any of the existing particle filters. This operation uses tiled mean-shift operators which begin by spanning the image plane, therby detecting all large peaks of motion, and spawns new particle filters when no existing filters are within a specified distance to each peak in the image plane. Particle filters are discontinued when particle tracks leave the image or when the total sum of all importance weights drops below a user-defined threshold. This, in addition to the fact that filters are discontinued when they begin to overlap heavily, helps to ensure that the filters do not mix into each other over time.

Fig. 3 shows the result of the particle filter tracking. Fig. 3a shows a situation in which three different particle filters have been spawned, each corresponding to a different object. Because the particles also maintain an estimate of each object's velocity, the system tends to be robust to objects whose paths cross with temporary occlusion of one of the objects by the other. Fig. 3b show the center of each particle filter—in this example all three moving objects are correctly identified (the large truck in the foreground did not move in the image sequence). Fig. 5 shows a shot of

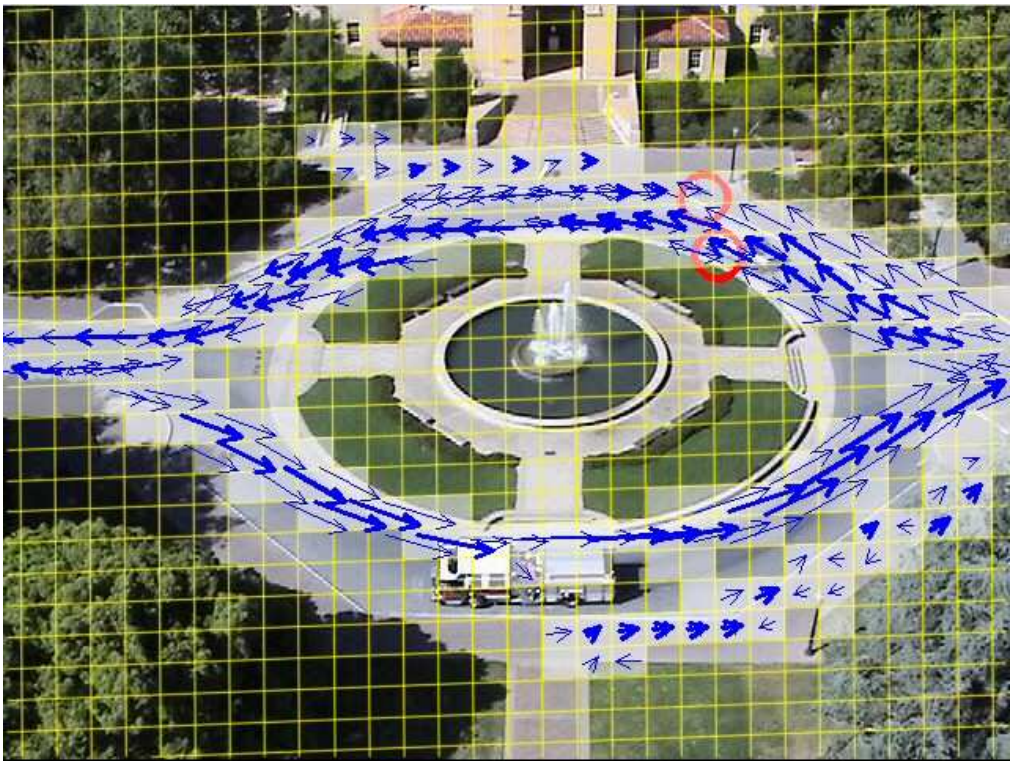


Fig. 4. Example of a learned activity map of an area on campus, using data acquired from a camera platform undergoing unknown motion. The arrows indicate the most likely motion direction modes in each grid cell; their lengths correspond to the most likely velocity of that mode, and the thickness represents the probability of motion. This diagram clearly shows the main traffic flows around the circular object; it also shows the flow of pedestrians that moved through the scene during learning. Video of learning grid being constructed over time is available at *www.motiontracking.info/learning-grid.avi*.

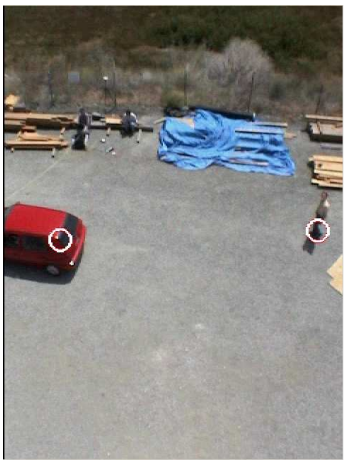


Fig. 5. Two moving objects being tracked in video taken from a helicopter as part of a DARPA demo.

tracking video taken from the Stanford helicopter during a demo for the Defense Advanced Research Projects Agency (DARPA). The two moving objects in the video have been correctly identified and tracked from overhead.

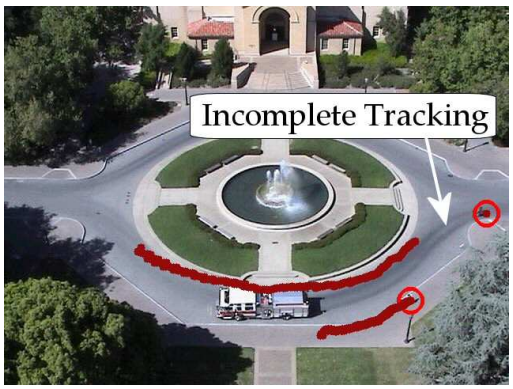
D. Learning The Activity-Based Ground Model

The final step of our approach involves the acquisition of the behavior model. For that, we anchor the map using features in the image plane that, with high likelihood, are not moving. In this way, the activity map refers to a projection of a patch of ground into the camera plane, even when that patch of ground is not presently observable by the camera. This ground plane projection remains static with respect to the ground and does not refer to relative locations in the camera image.

The activity map is then calculated by histogramming the various types of motion observed at different locations. More specifically, our approach learns a 4-dimensional histogram $h(x, y, v, \theta)$, indexed over x-y locations in the projection of the ground in the camera plane and the velocity of the objects observed at these locations, represented by a velocity magnitude v and an orientation of object motion θ. Specifically, each time the k-th particle filter's state $[s' \quad v']$, where $s' = \frac{1}{m} \cdot \sum_m [s_k^{[m]}]$ and $v' = \frac{1}{m} \cdot \sum_m [v_k^{[m]}]$, intersects with an x-y cell in the histogram, we increment the counter $h(x, y, v, \theta)$ where $x = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot s'$ corresponds to the x-coordinate of s', $y = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \cdot s'$ to its y-coordinate, $v = ||v'||_2$ to the magnitude of the velocity vector, and $\theta = \cos^{-1}[(v^{-1} \quad 0) \cdot v']$ to its orientation.

Fig. 4 shows the result of the learning step. Shown here is an activity map overlayed with one of the images acquired during tracking. Blue arrows correspond to the most likely motion modes for each grid cell in the projection of the ground in the camera plane; if the likelihood of zero motion

(a) Tracking without learned activity map



(b) Tracking with learned activity map

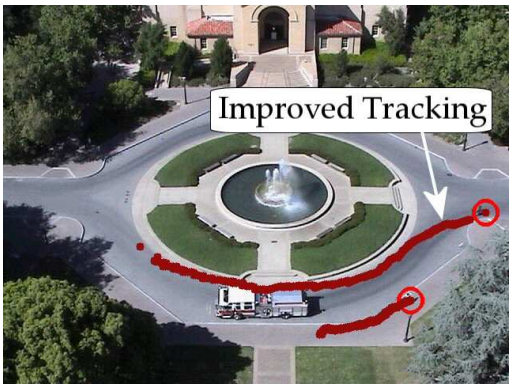


Fig. 6. Example of two tracks (a) without and (b) with the learned activity map. The track in (a) is incomplete and misses the moving object for a number of time steps. The activity map enables the tracker to track both objects more reliably. Full video is available at *www.motiontracking.info/comparison.avi*.

exceeds a threshold, no arrow is displayed. Further, the length of each arrow indicates the most likely orientation and velocity at each location, and its thickness corresponds to the frequency of motion. As this image clearly illustrates, the activity models acquired by our approach are informative of the motions that occur on the ground.

III. Applications

A. Using the Activity Model for Improved Tracking

To understand the accuracy of the activity model, we have applied it to improve the quality of the particle filter tracking. Specifically, in our improved tracking algorithm the importance weights $w^{[m]}$ are modified to take into account how well a feature's motion matches motion seen previously in that grid cell, according to the histogram h:

$$w^{[m]}_{improved} = w^{[m]} + k \cdot p(v^{[m]}_k, \theta^{[m]}_k \mid x^{[m]}_k, y^{[m]}_k) \quad (6)$$

The second term represents the probability of each particle's motion, given its $x-y$ location, times a constant scale factor k. This second term was added to and not multiplied by the original weights so that no single effect, either the original importance weight or the histogram-based motion reward, dominates.

Fig. 7. The single-frame alignment of two independent video sequences based on the activity-based models acquired from each. This registration is performed without image pixel information, only activity information from the learning grid.

In the rich literature of activity learning, we are unaware of an approach for using learned activity models for improving the accuracy of the motion tracker. On a 2100-frame test data sequence, tracking accuracy (defined as the number of correct tracks, minus the number of false positives, divided by the total number of moving objects) was 0.85 without using the learning data, and 0.89 when using the learning data. This corresponds to roughly a 27% reduction in the number of incorrectly identified or missed moving objects. Note: when a red indicator circle was located over a moving object it was counted as a correct track while when a red circle was located over a stationary part of the scene it was considered a false positive. No segmentation of the moving objects was performed once they had been identified. Automatic segmentation of moving objects has been addressed previously in the literature, and our goal for this paper was simply to recognize and track moving objects.

Fig. 6 compares the tracking without (top panel) and with (bottom panel) the learned activity map. More specifically, the top diagram is the result of using the standard importance weights to update the particle filters, whereas the bottom diagram uses our learned activity map for tracking, on independent testing data. As is easily seen, the track in the bottom diagram is more complete than the one in the top diagram, illustrating one of the benefits of our learned activity model.

B. Registration

By encoding the major activity modes of each learned grid cell as pixel intensity values, traditional image regis-

tration techniques can be applied to align video sequences based on observed activity (but not the actual image pixel values). In this manner, independent activity maps of the same terrain can be merged, and previously acquired activity maps can easily be updated with additional learning data. Furthermore, registration of activity-based models would enable autonomous systems to characterize and later identify locations on the ground based on the motions observed. For example, an autonomous helicopter or other aerial platform could distinguish a four-way stop from a traffic circle and orient itself based on the motions of the vehicles it observes. In a similar manner, video sequences of terrains lacking sufficient static image landmarks for traditional registration or whose characteristics change over time (e.g. a desert road or maritime shipping channels) could still be aligned.

Fig. 7 shows a single-frame alignment of two independent video sequences based solely on the activity-based models acquired from each video. While the alignment is not perfect, surprisingly accurate results can be obtained solely from the observed motion data.

IV. Discussion

We have presented a system for learning activity models of outdoor terrain from a moving aerial camera. Our approach acquires such models from a camera that is undergoing unknown motion. To identify moving objects on the ground, our approach combines image-based feature tracking with an EM-approach for estimating the image transformation caused by the camera's ego-motion. This identifies features whose motion is counter to the flow induced by the estimated ego-motion. We then apply multiple particle filters to identify and track moving objects. The object motion is then cached into a histogram that learns the probability distribution of different motions at different places in the world. Applications of the learned activity model include improved tracking and global registration of two different models based on the activity patterns.

Our approach runs at 20Hz on a 2.4GHz PC. To help distinguish slowly moving objects from the background, and increase the disparity between ego-motion and object motion, a full calculation is performed once every six frames. Particle filter information for interleaved frames is interpolated.

While most elements of our approach are well-known the literature, we believe that it defines the state-of-the-art in finding moving objects on the ground from a helicopter platform. Further, we believe that the use of learned activity models for tracking and registration is unique. Certainly, our system has been found to be robust in tracking moving objects and learning useful activity models of ground-based motion. These models have proven to be applicable to problems of general interest.

References

[1] G. Konecny. *Geoinformation: Remote Sensing, Photogrammetry and Geographical Information Systems*. Taylor & Francis, 2002.

[2] F. Lu and E. Milios. "Globally Consistent Range Scan Alignment for Environment Mapping." *Autonomous Robots*, 4:333–349, 1997.

[3] D. Hähnel, W. Burgard, and S. Thrun. "Learning Compact 3D Models of Indoor and Outdoor Environments with a Mobile Robot." *Robotics and Autonomous Systems*, 44(1), 2003.

[4] S.B. Williams, G. Dissanayake, and H. Durrant-Whyte. "An Efficient Approach to the Simultaneous Localisation and Mapping Problem." In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 406–411, Washington, DC, 2002.

[5] D. Ferguson, A. Morris, D. Hähnel, C. Baker, Z. Omohundro, C. Reverte, S. Thayer, W. Whittaker, W. Burgard, and S. Thrun. "An Autonomous Robotic System for Mapping Abandoned Mines." In S. Thrun, L. Saul, and B. Schölkopf, editors, *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. MIT Press, 2003.

[6] D. Makris and T. Ellis. "Automatic Learning of an Activity-Based Semantic Scene Model." *IEEE Conf. on Advanced Video and Signal Based Surveillance*, pages 183-190, 2003.

[7] C. Stauffer and W. Eric L. Grimson. "Learning Patterns of Activity Using Real-time Tracking." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:747–757, 2000.

[8] P.J. Burt, J.R. Bergen, R. Hingorani, R. Kolczynski, W.A. Lee, A. Leung, J. Lubin, and H. Shvayster. "Object tracking with a moving camera." *Proceedings of the Workshop on Visual Motion*, 2–12, 1989.

[9] B. Jung and G. Sukhatme. "Detecting Moving Objects using a Single Camera on a Mobile Robot in an Outdoor Environment." *8th Conference on Intelligent Autonomous Systems*, pages 980-987, 2004.

[10] J. Vermaak, A. Doucet, and P. Pérez . "Maintaining Multi-Modality Through Mixture Tracking." *Int. Conf. on Computer Vision*, pages 1110-1116, 2003.

[11] J. Shi and C. Tomasi. "Good Features to Track." *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 593–600, 1994.

[12] J. Bouguet. "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the Algorithm." *Intel Corporation, Microprocessor Research Labs* 2000. OpenCV Documents.