Change Detection From Multiple Camera Images Extended to Non-Stationary Cameras

Keith Primdahl,[1] Itai Katz,[2] Oren Feinstein,[3] Yi Lang Mok,[4] Hendrik Dahlkamp,[5]
David Stavens,[6] Michael Montemerlo,[7] and Sebastian Thrun[8]
Stanford University

Summary. We describe an approach for analysis of surveillance video taken from moving vehicles making repeated passes through a specific, well-defined corridor. Our goal is to detect stationary objects which have appeared in scenes along the established route. Our motivation is to address security concerns in hostile theaters where stationary surveillance cameras would be destroyed almost immediately; yet mobile camera platforms; i.e., group transport vehicles/convoys are plentiful. Challenges include illumination changes from different time/day, and handling parallax resulting from our non-stationary camera. We provide an example using artificial surveillance images taken on the Stanford University campus. Scale-up to critical security theaters would be straightforward. The approach is equally applicable to images collected by aircraft.

Key Words: image differencing, extended Kalman filter, object detection, parallax, segmentation, Sobel gradient, surveillance.

1 Background

For our surveillance application, we desire to establish a known environment; albeit, not necessarily mapped, and possibly with non-threatening changes over time. In addition, our environments are not suitable for fixed surveillance cameras due to hazards. For example, a roadway between safe zones in a hostile, possibly combative environment, where stationary surveillance cameras would be quickly destroyed by insurgents. Other possible applications include disaster scenes, natural or otherwise, where it is important to detect continuing structural failure while executing rescue operations. So, our challenge is to detect change in scenes, especially new or moved objects, in two or more videos taken sequentially from mobile platforms. Specifically, in sequential videos from different viewpoints, under varying time-of-day and lighting conditions.

Computer vision techniques applied to video from stationary cameras are widespread, both in the literature and in everyday use. One of the authors[9] participated in development of a traffic monitoring system where aging gray-scale cameras located along metropolitan interstate highways are analyzed for changes in the stationary background. With a moving-average background image, detection of new stationary

[1] Stanford University, Stanford, CA 94305; also with Lawrence Livermore National Laboratory, Livermore, CA 94550, primdahl@stanford.edu

[2] Stanford University, Stanford, CA 94305, itai@stanford.edu

[3] Stanford University, Stanford, CA 94305, feinstei@stanford.edu

[4] Stanford University, Stanford, CA 94305, yilangmok@stanford.edu

[5] Stanford University, Stanford, CA 94305, hendrik@dahlkamp.de

[6] Stanford University, Stanford, CA 94305, dstavens@stanford.edu

[7] Stanford University, Stanford, CA 94305, michael.montemerlo@stanford.edu

[8] Stanford University, Stanford, CA 94305, thrun@stanford.edu

[9] Itai Katz, while at University of Maryland.

objects (e.g., stranded motorists) is a straightforward application of background subtraction. Changes in illumination are incorporated into the background image by virtue of maintaining the moving average. Our work extends these applications to the more difficult case of non-stationary cameras under changing illumination.

The attentive reader will immediately recognize that non-stationary camera location will introduce parallax among objects in the scene. Parallax is most significant for objects nearest the camera; however, for a forward-facing camera on a forward-moving vehicle, all objects eventually approach the camera. Our studies purposely introduce a lateral shift in camera position, thereby resulting in significant parallax. Left unaddressed, parallax will make alignment of previously-present objects impossible, thereby leading to their detection as new objects. While our approach to mitigating parallax is simple, we have not found mention of it in the literature; in fact, very little is published regarding image analysis from non-stationary camera positions.

Our trials were successful in detecting new objects under overcast and bright-sunlight conditions. In addition, we performed a live demonstration at Stanford University, from a moving vehicle, with the scene image updating at approximately six Hz.

Finally, it is important to recognize that false positive detections cannot be tolerated in the described scenarios. Application of our work has the potential to generate a large amount of data in critical surveillance environments. Even a seemingly small rate of false positives would quickly overwhelm any security staff, and thereby negate the value of collected data.

2 Related Work

A wide range of approaches for detection and tracking of stationary and moving objects from stationary cameras is provided in the literature. Considerably less attention has directed towards non-stationary cameras. Structure from motion approaches [1][16], including extensions of Simultaneous Localization and Mapping (SLAM) [6] address the important application of a robot learning (mapping) a new environment. Along with Szeliski, Levin [3] presents a map correlation study for the case where camera positions are unknown. Here, they extend SLAM by matching features on Red-Green-Blue (RGB) histograms, moments, and Harris corners. Features are matched via loopy belief propagation. While SLAM is implemented with a moving camera, change detection would be a significant extension.

Sand and Teller [7] describe a method for alignment of video images where some change between scenes is tolerated. They first determine a pixel matching probability using 3x3 maximum and minimum-intensity filters. Dissimilarity values are determined over a range of possible uniform offsets, with scores averaged over three channels. For each candidate match, a motion consistency probability is calculated by starting with a Harris corner detector. Locally-weighted linear regression is used to find a fit over nearby pixels, where “nearby” is an adaptively-sized Gaussian filter. Each fit is subjected to Kanade-Lucas-Tomasi (KLT) fitting, while those that exceed a threshold are refined by performing another iteration. Finally, the product of pixel matching probability and motion consistency probability determines the best offset. The mapping is interpolated and extrapolated to obtain and apply a non-uniform, dense correspondence field. While their *Video Matching* approach addresses some of our concerns, Sand and Tellers objective is more generally to ignore change rather than detect and highlight change. There is no mention of parallax.

Szeliski [10] and Szeliski with Shum [11] provide a review of relevant transforms, with direct and feature-based alignment techniques for determining correspondence in overlapping images. They emphasize rough-alignment via feature-based methods, followed by determination of more accurate correspondence using patch-based alignment with a robust error metric. Ghosting due to parallax is handled only locally, and after alignment.

Scharstein, et.al. [8] provide a taxonomy of dense two-frame stereo-matching algorithms. However, stereo matching requires parallax rather than seeks to eliminate it.

Toyama, et.al. [15] provide an approach to background maintenance they title “Wallflower.” Addressing surveillance applications, they identify a suite of difficulties, many of which are similar to those faced in our efforts; e.g., changing time of day, changing illumination, shadows, waving trees, and camouflage. They develop a three-level approach: 1) pixel-level, for prediction of the expected background, 2) region-level for component fills in homogenous regions, and 3) frame-level, for global changes. Because the background described here is far from the observer, camera position and parallax is disregarded.

Lowe’s classical paper [4] describes a Scale Invariant Feature Detector (SIFT). His approach provides good localization, while performing under changes in scale, rotation, illumination, and some cases of deformation. For each feature, Lowe computes a feature signature by developing a local coordinate system for each, characterizing each feature independent of scale, eliminating edges (by substituting ratio of eigenvalues), normalizing for brightness, and thresholding for camera saturation. A catalog of features is thus maintained. While SIFT clearly has application to alignment of images, and its “invariant” nature might be helpful in aligning images from differing camera positions, his work stops short of addressing the necessary transformations for alignment.

Shi and Tomasi [9] present an extension to classical Lucas and Kanade, introducing a measure of feature dissimilarity. They provide evidence that linear warping and translation are adequate for image motion when measuring dissimilarity. Specifically, they begin with conventional feature detection, then use Lucas-Kanade to track pure translation. Affine warping provides registration for tracking general motion. New tracks are started as necessary. Here also, the work is not directed towards transformations necessary for alignment of images from different camera positions.

In summary, we are unable to find where the salient elements of our work has been previously addressed in the literature; specifically, object detection by removal of parallax, alignment of image from differing camera positions by appropriate transformations, and robustness under varying illumination.

3 Approach

Unique to all possible approaches for object detection is the need to first pair images from different videos. Our images are obtained with position information by virtue of the camera’s extended Kalman filter (EKF) [6]. The pose data further combines measurements from global positioning system (GPS), differential GPS, GPS compass, wheel odometry, three gyroscopes, and three accelerometers; thereby providing output in six degrees of freedom (x, y, z, roll, pitch, yaw).

Alignment of one or more images is a recurring challenge in computer vision. Well-known approaches are often subdivided into *dense* and *feature based*. Dense approaches generally establish a similarity metric, which is applied on a pixel-by-pixel basis. Well-known approaches include disparity matching and the Lucas-Kanade, with zero optical flow corresponding to perfectly-aligned images. Feature-based methods include Canny Edge Detection, Harris Corner Detection, and the previously-mentioned SIFT approach. After pairing collected images, we reverse perspective projections, define regions of interest, then translate and rotate to a common image coordinate system in accordance with the collected pose data. This approach eliminates parallax, as we reverse the projection to a top view, where all objects and the road's surface are essentially the same distance from the camera. Incremental refinement of alignment is achieved by best-fitting Fourier transforms.

Following alignment, each frame pair is compared by subdividing the regions of interest into a grid, converting the images from RGB to Hue-Saturation-Value (HSV), then analyzing gradient on the Value (brightness) channel. The presence of new objects within each cell is detected by comparison of pixel counts for gradient above a threshold.

The feature-based methods mentioned above were generally developed to determine optical flow, yet could be applied to object detection as well. Perhaps less well-known is the Sobel gradient approach, where edges are determined from the first or higher partial derivatives in one or both directions [12]. Our trials confirmed that, of these methods, effectiveness of Sobel gradient detection was least affected by changed illumination between video image pairs. While similar to edge detection, Sobel gradient detection is a *dense* approach; specifically, we do not use any feature detection.

3.1 Pairing

The general approach to all our studies was to first record one or more "base" videos, without objects, along a short stretch of "off-road" at Stanford University. Next, objects of varying color and texture were placed along the roadway, within a six-meter wide "lane" of interest. Subsequent videos were taken with the vehicle proceeding along the same trajectory, or approximately three meters to each side of the base trajectory. Weaving was included in some trials, with vehicle yaw of approximately ±ten degrees. Our [moving] region of interest is a six-meter square, projected in front of the vehicle, and outlined in green on the image below.

GPS and inertial-based pose estimation provide an estimate of each image's global registration. These estimates are fused using a Sigma Point Kalman Filter (SPKF) [5]. The sigma-point filter offers an important benefit over the more traditional Extended Kalman Filter (EKF) [17]. When faced with a nonlinear system, the EKF approximates the predictions as the prior mean (no expectation!) and the covariances by linearization. Such approximations can be arbitrarily inaccurate. The sigma-point filter differs from the EKF primarily in how the first and second order posterior is calculated. In the sigma-point filter, the Gaussian state is represented by a weighted set of samples that accurately depict the mean and covariance of the prior. Now arbitrary nonlinearities are captured to the second order. Nonlinearities in the exponential family are captured to the third.

For our tests we used the Stanford Roadrunner Robot. The Roadrunner utilizes GPS with satellite-based differential corrections, a GPS compass, a six degree of freedom inertial measurement unit, and wheel odometry.

Frames from the current and previously-recoded videos are paired based on nearest Euclidian distance. We found resolution to be approximately ¼ meter when our GPS was first initialized. Based on our observations, along with those of Stanford colleagues, we concluded that the GPS resolution improves over the first two hours or so, sometimes achieving repeatability of a few centimeters. For these studies roll and pitch of the vehicle were ignored. Accounting for roll and pitch would be straightforward, and probably warranted for undulating terrain.

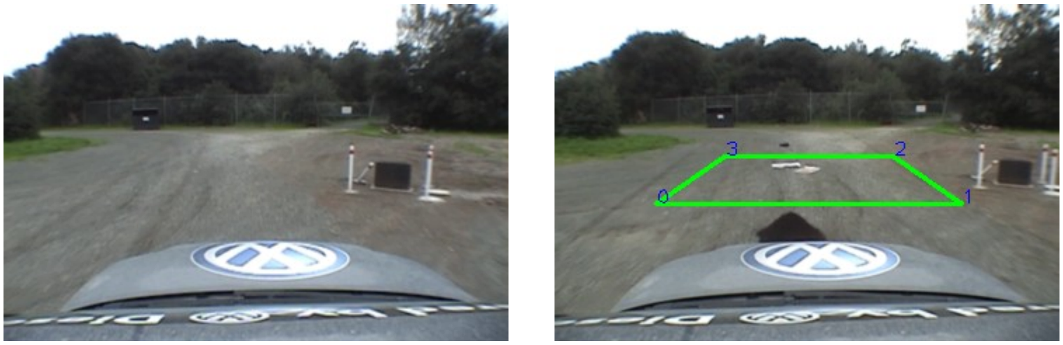

Fig. 1. Video frames as acquired. Left: Without Objects.
Right: With objects and highlighted 6m x 6m region of interest.

3.2 Reverse Perspective Projection and Alignment

Early in our studies, we recognized that the road's surface could be assumed planar. For any single plane surface in a perspective projection, it is possible to perform a reverse projection which transforms the selected plane to an orthonormal view; i.e., a top-down orthonormal view in our case. Of course, other planes of the perspective projection, and non-planar objects, will be distorted by this transformation, often severely distorted. For our efforts, plane surfaces other than the roadway can be ignored. While objects on the roadway are not strictly in the same plane as the roadway, our goal is sufficiently accurate alignment to prevent inadvertent detection as new objects. Accordingly, some distortion of these objects can be tolerated.

Our transformation of both images to a top view could be accomplished by a mere rotation about a single [known] camera axis. However, we employed an approach suggested by Hartley and Zisserman [2]. Selection of any four points on the plane of interest provide eight correspondences. Since the required transformation matrix requires nine parameters, only a scale factor is left undefined. We leave element [3,3] of the transformation matrix equal to one, while incorporating scale into our point correspondences. Specifically, the four image points of the perspective view are selected to correspond to a square in the top-down orthonormal view. The corresponding points of the square are selected for a reasonable fill of the top-down orthonormal image. The determined points for both the perspective and top-down orthonormal view are centered about the point where the optical axis meets the road. The calculated translation matrix is constant for a particular camera focal length, height, and angle—generally fixed values once a camera is mounted to a vehicle.

Once we have both images in top-down orthonormal views, the optical axes are parallel. For each frame pair, we use the GPS coordinates, similar triangles, and the camera's focal length to determine the number of pixels to translate our "base" image in image X and Y coordinates (our "base" image is "without objects" below). Similarly, we rotate our "base" image about the camera axis to correct for difference in yaw between the two camera poses. For execution, we combine the translation and rotation into a single transformation matrix. Unlike the transformation to reverse the perspective projection, this translation/rotation matrix must be calculated by our code for each frame pair.

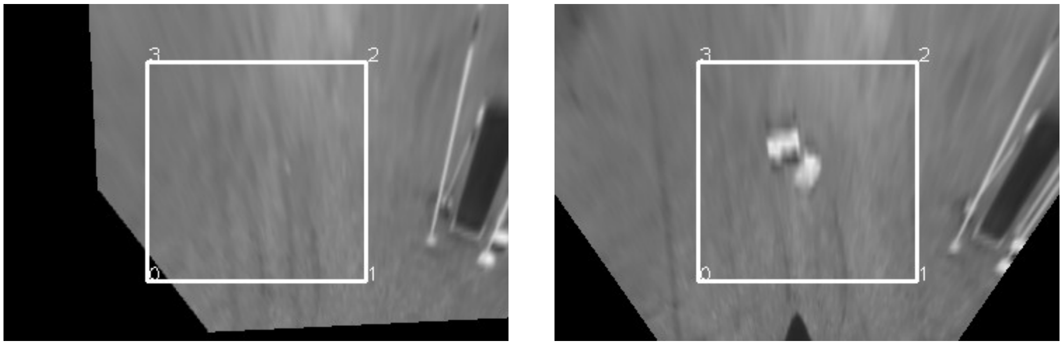

Fig. 2. Transformations to Top-Down Views, both with highlighted 6m x 6m Regions of Interest. Left: Without objects, and aligned to the image with objects. Right: With objects.

Following the combined translation and rotation to correct for pose information, we refine the alignment by performing a Fourier transform on both images, taking the product, then reversing the transform. Szeliski [10] suggests that pure translation, as is the case here, is the only place where this Fourier-based convolution is effective.

3.3 Segmentation and Gradient Detection

For our change detection algorithm, we convert the two images from RGB to HSV, and extract the Value (brightness) channel. A first-derivative Sobel gradient filter was applied to the Value channel, with an empirically-determined threshold. Pixels above the threshold are set to "white," and can be compared between images.

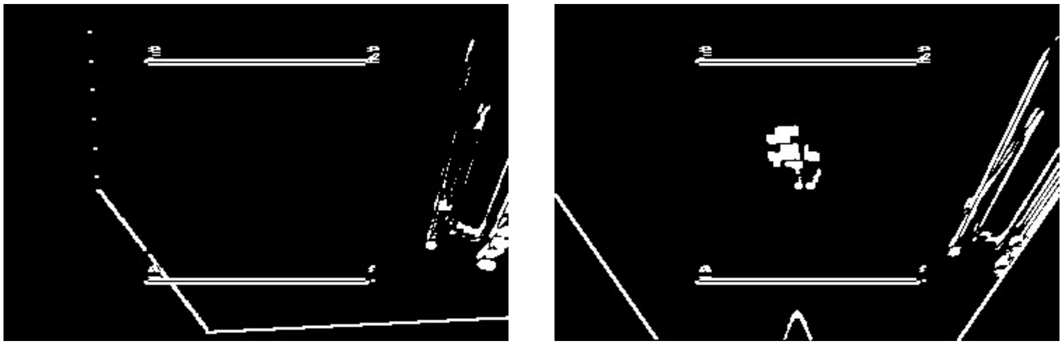

Fig. 3. Sobel Gradient Images, with white pixels for above threshold. Left: Without Objects. Right: With objects.

A direct comparison of the two images would require pixel-for-pixel alignment. While the alignment methods described above have provided good results, even small translation errors result in false positives. To relax the alignment restriction, we overlay a grid on the 6m x 6m region of interest, and we compare pixel counts within corresponding grid cells. This segmentation allows alignment error of approximately one-half of the grid cell width, at a loss of some object localization. If the number of white pixels in corresponding cells differs by a second, empirically-determined threshold, that cell is designated as identifying a new object.

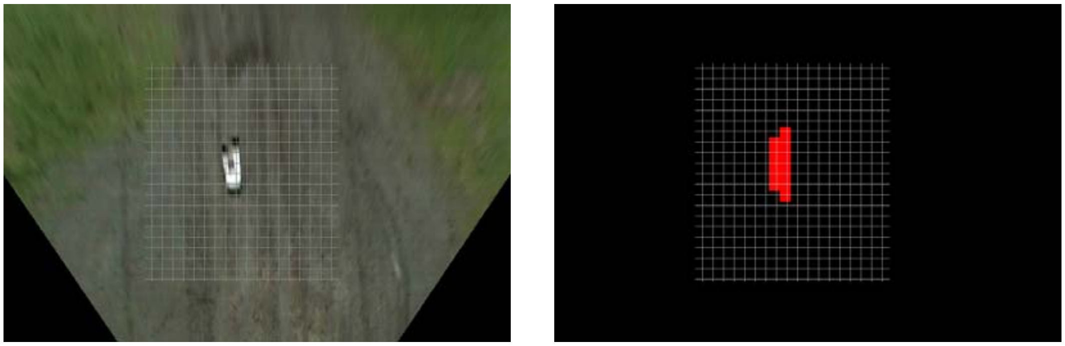

Fig. 4. Left: grid overlay on top-down view with object.
Right: grid with red cells where object has been detected.

Attempts to detect objects via Canny Edges on gray-scale images, and Sobel gradients of Hue and Saturation were found to be less effective than taking the gradient of the Value channel. In the case of Canny Edges, the detector missed many natural objects that lacked well-defined edges (e.g., loose piles of potting soil or small objects). The gradients of Hue and Saturation channels yielded many false positives because these are inherently noisy; in particular, Hue changes frequently, even within the same object. For computational economy, we chose a Sobel gradient in the vertical direction only; the typical motion flow in our trials yielded greater vertical gradients than horizontal gradients.

3.4 Visualization of New Objects

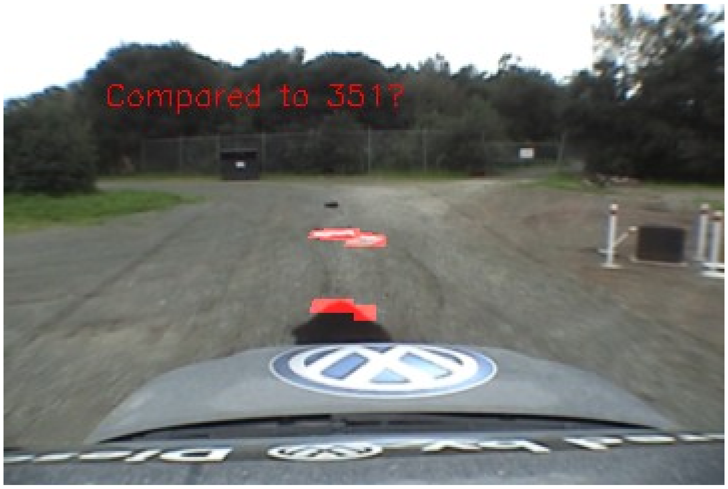

Fig. 5. Red cells indicating a detected object inside the region of interest are blended with the as-acquired video images. More distant objects will be similarly highlighted as the region of interest overtakes their locations.

For visualization of detected objects, we color grid cells to red, and then perform a reverse projection to convert from the orthonormal view back to the original perspective—using the inverse of the constant-value transformation matrix described earlier. The warped grid is overlaid atop the original image, yielding a live view of detected objects.

4 Results and Conclusions

We have successfully identified new objects out of videos taken from differing camera locations. Repeated trials of straight trajectories, with camera location displaced up to three meters laterally were repeatedly able to detect new objects within the 6m x 6m region of image (perhaps more accurately described as a six meter-wide traffic lane of interest). Slightly winding trajectories, with yaw varying ± 10 degrees were similarly successful.

While most of our work was under overcast conditions, a live demonstration was held under [unexpected] bright sunlight and blue sky, with well-defined shadows in our field or view. Some preliminary false positives were eliminated via minor adjustment to the Sobel gradient threshold. Thereafter, our live demonstration was successful; duplicating our results under overcast conditions. Our demonstration was performed in a moving vehicle, with the scene image updating at approximately six Hz.

5 Future Work

We have identified several possible areas for future investigation and improvement:

- Perform extensive data acquisition over varying and carefully recorded conditions; e.g., to optimize thresholds/parameters.
- Correct for roll and pitch of vehicle. Videos made from our of top-down orthonormal-views show a minor non-uniform oscillation due to varying roll and pitch of the vehicle. We can easily account for pitch and roll by including a correction to both top-down orthonormal views; specifically, make minor rotations to correct for roll and pitch of the vehicle, as provided by the camera pose information. The resulting top-down orthonormal images will then have a stable, vertical optical axis.
- Relax our assumption of a planar road; i.e., accommodate an undulating road. As described above, we assume the road is planar not only over the 6m x 6m region of interest, but that the same plane also extends under the vehicle (this assumption is reflected in our constant transformation matrix used to reverse the perspective projections). An undulating road undoubtedly violates this assumption. To improve on our assumption of a planar road, we can determine a best-fit plane through point cloud data from one or more laser range finders. For a severely undulating road, it may be possible to facet the region of interest, perform separate transformations to reverse the perspective projections, then combine the facets into a single top-down orthonormal view of a single plane.
- Explore feature-matching approaches to detect changes based on texture. Combine texture-based results with brightness-based results described here. Weight these results, perhaps adaptively, to increase sensitivity to new objects without false positives.
- Include normalization of brightness. Our initial studies included Canny edge detection, where we found significant improvement in detection when images were normalized for brightness and contrast [7]. We concluded that variable illumination is a significant factor; however, can be mitigated by normalization. Still, our success with Sobel gradient detection—without normalization—was

immediate. We intend to study the possible benefit of similarly normalizing the Value channel prior to Sobel gradient processing. We postulate that normalization will allow us to lower the Sobel gradient threshold, thereby increasing the Sobel gradient sensitivity without introducing new false positives.

- Investigate context-based recognition. Going beyond mere object detection, Torralba, et.al. [14] describe an approach for recognition of what an object is, using context-based information; that is, using recognition and knowledge of surroundings to recognize specific types of objects (e.g., a coffee maker being easier to recognize in a kitchen, rather than against a blank background). It may be possible to increase the sensitivity of our approach, without generation of false positives, by including context-based recognition of objects; i.e., objects to be ignored.

6 Acknowledgements

The authors gratefully acknowledge financial support from the Stanford Racing Team.

References

1. D.A. Forsyth and J. Ponce. *Computer Vision, A Modern Approach*. Prentice Hall, Upper Saddle River, NJ. 2003.
2. R. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, Cambridge, UK, Sept. 2003.
3. A. Levin and R. Szeliski. “Visual odometry and map correlation.” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2004)*, volume I, pages 611-618, Washington, DC, June 2004.
4. D.G. Lowe. “Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, volume 60, issue 2, pp 99-110, Nov 2004.
5. Rudolph van der Merwe and Eric A. Wan. "Sigma-Point Kalman Filters for Integrated Navigation." In *Proceedings of the 60th Annual Meeting of The Institute of Navigation (ION),* Dayton, OH, Jun, 2004.
6. M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. “FastSLAM: A factored solution to the simultaneous localization and mapping problem.” *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 593-598. AAAI Press, 2002.
7. P. Sand and S. Teller. “Video Matching.” *ACM Transactions on Graphics (TOG)* 22, 3, 592-599. NEED YEAR
8. D. Scharstein, R. Szeliski, and R. Zabih. “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.” *In Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision*, Kauai, HI, Dec. 2001.
9. J. Shi and C. Tomasi. “Good features to track.” In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'94), pp 593-600, IEEE Computer Society, Seattle, Washington, June 1994.
10. R. Szeliski. *Image alignment and stitching: A Tutorial.* Pre-print provided by author. Microsoft Corporation, Redmond, WA.
11. R. Szeliski and H.Y. Shum. “Creating full view panoramic image mosaics and texture-mapped models.” *Computer Graphics (SIGGRAPH'97 Proceedings),* pages 251-258, August 1997.
12. S. Thrun. Course notes. *CS223B, Introduction to Computer Vision.* Stanford University, Stanford, CA. 2005.

13. S. Thrun. “Robotic Mapping: A survey.” In *Exploring Artificial Intelligence in the New Millennium*, pp 1-36. Morgan Kaufmann, 2002.
14. A. Torralba, K. Murphy, W. Freeman, and M. Rubin. “Context-Based Vision System for Place and Object Recognition.” *IEEE Intl. Conference on Computer Vision (ICCV)*, Nice, France, October, 2003.
15. K. Toyama, J. Krumm, B. Brummit, B. Meyers. “Wallflower: Principles and practice of background maintenance.” *International Conference on Computer Vision,* September 1999, Corfu, Greece.
16. E. Trucco and A. Verri. *Introductory techniques for 3-D Computer Vision*. Prentice Hall, Upper Saddle River, NJ, 1998.
17. Eric A. Wan and Rudolph van der Merwe. "The Unscented Kalman Filter for Nonlinear Estimation." In *Proceedings of Symposium 2000 on Adaptive Systems for Signal Processing, Communication and Control (AS-SPCC), IEEE,* Lake Louise, Alberta, Canada, Oct, 2000.