

Localization for Lost Robots through Scene Recognition (May 2008)

Pedro Davalos, *Member, IEEE*

Abstract—Mobile robot localization is a common problem in the robotics community, yet many proposed solutions are hardware dependent, requiring various sensors with prior environment information, or solutions that require continuous feature tracking such as the simultaneous localization and mapping (SLAM) framework, where prior motion dynamics or accurate odometry is required. However, these localization techniques are not robust to handle localization for single-camera lost robots, where the system can encounter temporary lack of features to track, thus hindering the system lost without any method to recover and regain correspondence to the previously built map. In this paper, we present a technique that addresses these issues and achieves robust performance in localizing a lost robot through finding correspondence between the current scene and a previously built map.

Index Terms—Localization, Lost Robot, Mapping, Scene Recognition, SIFT

I. INTRODUCTION

THIS document presents a practical near real-time solution for localization of lost robots. Localization is an essential part of robotics for situational awareness and for navigation. Localization is concerned with continuously knowing the current position relative to a fixed reference frame, such as the starting point, and knowing the position relative to multiple fixed landmarks, such as prominent fixed objects in the environment.

The ideal solution would consist of a robust system that can achieve continuous localization to navigate in an unknown environment with unknown motion dynamics or odometry without active sensing, utilizing exclusively a single camera as a sensing device, while handling temporary blindfold displacement, temporary motion with lack of distinctive features such as from a blank wall, clear sky, or bright light saturation. These situations of motion without features are common in practical applications, and they have not been addressed by current robot localization research that focuses on feature tracking. Furthermore, an ideal localization system would allow a robot to start without any prior knowledge, but

it would also allow initialization with a prior map of the environment, where the robot would have to localize with respect to the landmarks on the map. This feature provides significant capability for map expansion during missions where continuous operation is impossible. This capability to initialize with a known map also proves invaluable as it allows navigation for robots that may shift position when on standby or when disabled.

These factors contributed to our motivation for investigating and designing a robust and versatile system capable of handling localization for a lost robot, while at the same time still performing accurate continuous real-time localization when current location relative to the map is available, and when distinctive features for tracking are also available.

This presents a dual mode of operation, where the first case involves normal tracking mode, with known location and sufficient features to track, and the second case is the lost robot mode, where we have a map of the environment but our location is unknown. This paper focuses on the second mode of operation, dealing with the lost robot localization, but we also discuss potential integration issues between the two modes of operation.

Additional motivation for this work includes our constraint on utilizing only a single camera as a sensing device, as this requirement allows for an extremely versatile system capable of easy deployment and integration onto any robotic platform. This versatility is primarily due to the lack of odometry data, where localization systems can be deployed as wearable sensors or as payloads on non-robotic platforms for auxiliary and enhanced navigation systems.

II. RELATED WORK

Multiple Localization techniques have been previously developed and successfully demonstrated, however most of these methods apply only to special cases with specific hardware or environment constraints, or yet some methods provide a solution as a batch process for offline post-processing of the data.

Some of the early work on Localization was conducted by Smith [1], Leonard [2], and Manyika [3], where they utilized the Extended Kalman Filter (EKF) as a particle filter for updating and predicting the state estimate, which allowed for the theoretical groundwork of the Simultaneous Localization and Mapping (SLAM) framework. Yet their solutions were

Manuscript received May 7, 2008. This work was supported in part by the Departments of Computer Science at Texas A&M University in conjunction with the Spacecraft Technology Center of the Texas Engineering Experiment Station (TEES).

P. Davalos is with Texas A&M University, College Station, TX 77843 USA (phone: 979-845-8768; e-mail: pedro@tamu.edu).

designed for active sensing, and their theory lacked the processing power for practical implementation at their time. Subsequent work was conducted by Csorba [4], where by the late 1990's sensor measurements were becoming more reliable and less noisy, which allowed for optimism and popularity of the SLAM framework.

Further significant contributions have been presented by Thrun [5], and Montemerlo [6], where they developed the GraphSLAM and FastSLAM algorithms that rely on lidar or SICK laser range finder sensors, which were demonstrated to have positive results, however, these techniques are designed for offline batch processing, making them unable to perform in real time applications.

To address most of these issues, Davison [7],[8] presents a novel approach to the localization problem by using a single camera system, MonoSLAM, without odometry, capable of continuously localizing in real time. As this is a vision based system, Davison utilized Shi and Tomasi's method [9] to find distinctive features in the scene, as this provides an efficient method to populate the sparse map of visual features. However, two drawbacks from this approach include the limitation of the number of landmarks due to the quadratic growth of the covariance matrix for the EKF, and the second drawback from MonoSLAM is the inability to recover from a temporary loss of tracked features, when in motion without odometry.

Subsequent work by Davison and Clemente [10] has addressed the first limitation of MonoSLAM, where the landmark limit has been removed by the use of multiple overlapping maps, with each map containing a fixed number of landmarks. Furthermore, Montiel and Civera [11] have also demonstrated an enhancement to the original MonoSLAM algorithm where initialization is automatic without the need for calibrated initialization, since features are detected through an inverse mechanism for 3d correspondence.

Similar research of Monocular SLAM algorithms has been conducted by Smith [12], where successful results illustrate the feasibility of utilizing straight lines as landmarks instead of feature points. As straight lines have the potential of increased robustness of viewpoint and lighting invariance.

Solving vision based lost robot localization has been previously attempted through object recognition [13] and Scene Recognition [14]. Se [13] attempted to solve the full localization problem through Scale Invariant Feature Transform (SIFT) Points, utilizing Lowe's method [15],[16], for extracting and representing distinctive high-quality landmarks. Se's method for localization is adequate for small indoor environments with short term motion, as longer increments could introduce drift to the localization estimate. Se's approach also required a trinocular stereo system for depth estimation, thus the small environment constraint, as stereo disparity fails beyond the calibrated limit. SIFT points have also been successfully used for object detection in [15],[17], as SIFT has been widely demonstrated successfully for multiple computer vision applications. Recent extensions

of the basic SIFT algorithm have been developed by Ke and Sukthankar [19] by applying Principal Component Analysis (PCA) to the SIFT feature descriptor, which demonstrate a significant improvement in computational performance and improvements in object recognition performance, making PCA-SIFT an interesting alternative to the original SIFT descriptors for any of the multiple applications using SIFT, including scene recognition for localization.

An interesting alternative to localization is through scene detection [14],[18] where Oliva and Torralba successfully demonstrate the capability of classifying and recognizing scenes through an elegant representation based on spectral content in the scene. Spectral information from scenes has been shown as an effective method for scene representation and recognition, with similarities to the cognitive process of scene recognition and localization by human perception [14].

Further recent work on Localization and Mapping has been conducted by Bowling, Wilkinson, and Godsi [20], where they introduce the different perspective of "Subjective Mapping" by Action Respecting Embedding (ARE) [21]. Which is a framework based on a probabilistic approach that incorporates past experiences from integrating observations and actions.

III. LOST ROBOT LOCALIZATION

Our goal of localizing a lost robot, in real time, with a single camera system, without any odometry or additional sensors, while at the same time allowing accurate positioning when operating in tracking mode presents unique challenges with various options to approach the problem.

Lost robot localization is defined for this work as the task to find the position of a robot, with a map of the environment where the current location is unknown, with a uniform probability distribution of position in the map. This scenario is not the case while tracking features in MonoSLAM [7] as the relative position is always known and constantly tracked during normal operation. Consequently, we must modify the MonoSLAM approach to incorporate a process for localization during lost-robot mode. Where this lost-robot mode persists until a position is established with reasonable uncertainty. Once a position with respect to the map is identified, normal MonoSLAM tracking mode can resume. Alternately, we could also modify the Inverse Depth [11] MonoSLAM approach by allowing an efficient real-time process for searching similar maps and merging maps once the corresponding relationship is established, as a new map would be started when initializing in lost mode. However, map similarity searches are unfeasible for the original MonoSLAM approach, as the map is a sparse collection of feature points, where the landmarks are not selected in a deterministic process. Furthermore, landmark points are saved and represented by their raw template.

To overcome this issue of stochastic landmark selection, we will assume that the MonoSLAM landmarks are represented

by SIFT Descriptors [15], where we can then process the full frame when in lost-robot mode, extract SIFT features, and compare the extracted features to the database, where only a subset of features need to be found in order to declare a successful match.

The process to localize a lost robot involves the assumption of an existing map, represented by a sparse collection of high quality feature points saved as SIFT descriptors with their corresponding scene label. Then the lost mode operation processes each incoming frame, extracts SIFT features, and searches the map for similar descriptors until a successful match is found. After a successful match is found, normal MonoSLAM operation resumes tracking features in continuous mode.

As an enhancement to our approach, the same process for lost robot localization can be applied while reducing the dimensionality of the SIFT descriptors. Each raw SIFT descriptor is represented by a 128-dimensional feature vector. This high-dimensional representation poses computational problems for storage and for nearest neighbor similarity searches, as expected from the curse of dimensionality. This step of dimensionality reduction emerged as an optimization to the similarity searches when computing scene correspondence.

Although dimensionality reduction is a research topic on its own, two popular techniques include Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). However, as we seek to preserve relative distances in the feature space, PCA is the method providing a sound theoretical solution since PCA captures information contained in the variance of the data. Consequently, through projections from the basis with the highest variance, we preserve relative distance and discriminating information while reducing the dimensionality by approximately an order of magnitude. Further analysis of the SIFT descriptor and dimensionality reduction through PCA is described in detail in the following section.

IV. SIFT DESCRIPTOR ANALYSIS

The SIFT Feature Point Descriptor is a 128-dimensional vector that describes each prominent landmark on input scenes. SIFT feature extraction is a common process, originally presented by Lowe [15], where the input scene is processed through a series of Gaussian Filters to find distinctive local extrema, which is a computationally expensive process, but an extremely effective method for representing landmarks due to the resulting invariance to scale, lighting, and viewing angle. The resulting descriptors, although configurable through the SIFT settings, have a standard dimension of 128, where each vector is normalized to unit length.

Our initial baseline test data consists of 10 indoor scenes as shown on Fig. 1, where each of the 10 input images is an 8-bit grayscale image with 120x160 pixels. SIFT feature extraction



Fig. 1. Baseline Data: 10 Input images corresponding to the map for lost robot localization. Each image is 8-bit grayscale with 120x160 pixels.

is a deterministic process which takes constant computational time (for a given frame size), although processing time varies in practice depending on the content of each scene (uniform frames with no distinctive patterns execute faster than elaborate complex content).

This baseline dataset generated a total of 889 SIFT descriptors, which is about 89 SIFT features per image, analyzing these 889 vectors through Principal Component Analysis results from solving the EigenValue problem of the Covariance matrix of the data. The resulting EigenValues are illustrated on Fig. 2, where the top 20 Principal Components capture about 70% of the variance. However, testing will also include the effect from changing the number of Principal

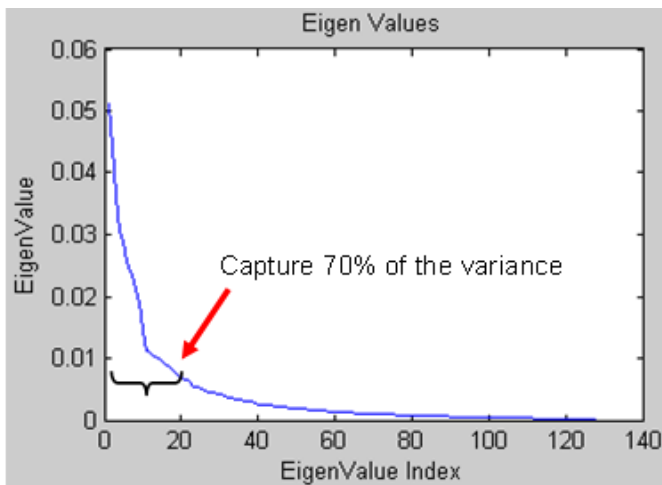


Fig. 2. Principal Component Analysis (PCA): EigenValues in descending order from the EigenVector solution of the Covariance Matrix of the data from the 889 vectors from the 10 original baseline image database.

Components, as the top 12 capture about 58% of the variance.

V. RESULTS

Our localization method for lost mode operation can be evaluated by analyzing performance results under various categories such as timing analysis, and localization performance by false positive rate and false negative rates. Where the false positives occur when a match is identified and a location established when in fact the robot is at a different position, and false negatives occur when the robot is at a known location with a familiar scene in the field of view, yet the robot fails to recognize the scene thus failing to establish its position.

All timing analysis experiments were performed on an Intel based PC with a 2.6 GHz Core2 Duo CPU and 4 GB of memory. Images were acquired using a Unibrain Fire-i 1394-Firewire Camera.

A. Timing Analysis: Baseline Database (10 images)

The goal of analyzing execution time of our lost mode localization algorithm is to verify feasible real-time operation. The main components of our algorithm involve SIFT feature extraction and similarity search, which is a nearest neighbor approach as we must find the closest match from the current scene to the map. Fig. 3 illustrates performance in lost mode when using RAW 128 dimensional SIFT features. The experiment was conducted by processing about 200 scenes not in the map/database. Fig. 4 illustrates the performance when using SIFT features with reduced dimensionality through PCA using the first 20 principal components.

Processing time is heavily influenced by the SIFT feature extraction which takes on average about 0.4 seconds. These experiments were conducted while searching for similarities on a database of 10 images. Each of the images in the database were represented on average by 89 SIFT features per image. Therefore the total database size consisted of 889 SIFT features.

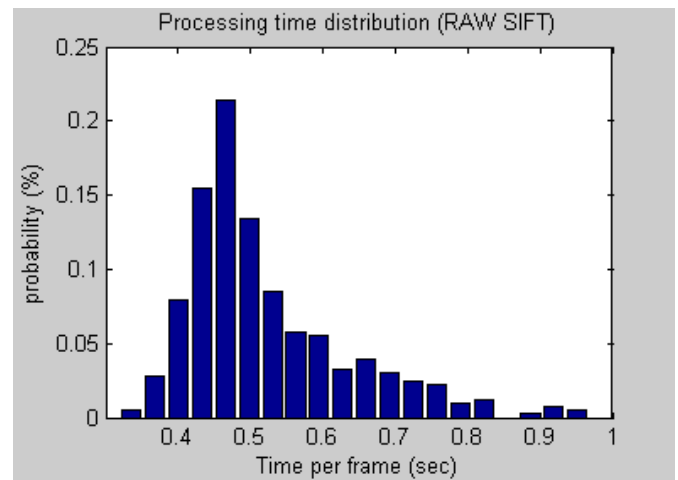


Fig. 3. Timing Analysis (10 images in database): Probability Distribution of execution time for each frame using RAW 128 dimensional SIFT features. Results from processing about 200 frames.

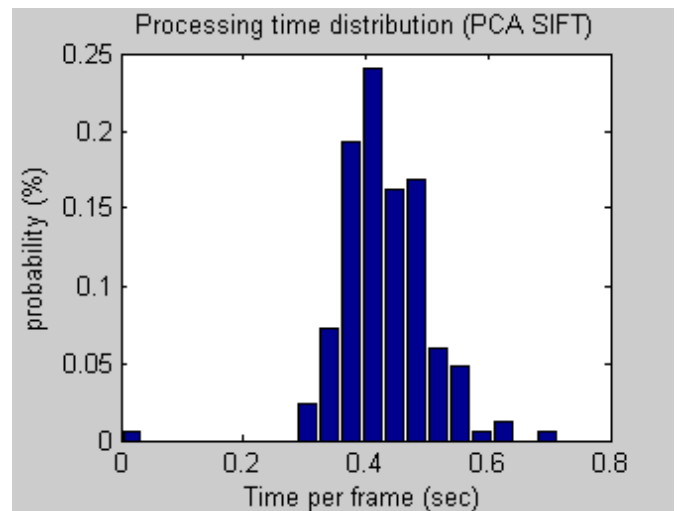


Fig. 4. Timing Analysis (10 images in database): Probability Distribution of execution time for each frame using SIFT features with reduced dimensionality through the first 20 Principal Components using PCA. Results from processing about 200 frames.

B. Timing Analysis: Extended Database (20 images)

This second experiment was conducted with a similar fashion as the previous experiment, except we used a database with twice as many images. This new map size of 20 images is capable of covering twice the area, which is sufficient to represent most indoor office areas with multiple rooms in the map.

As previously noted, the processing time to extract SIFT features is constant for each frame at about 0.4 seconds. The difference in processing time from the earlier experiment is due to the similarity search, which is now searching through 20 images, represented by 1778 SIFT features.

As we can observe from the results, processing time when using the PCA reduced dimensional SIFT features, we obtain a significant enhancement in performance, as the mean value from the RAW SIFT is about 0.75 seconds and the mean value from PCA-SIFT is about 0.45 seconds. Fig. 5a illustrates the

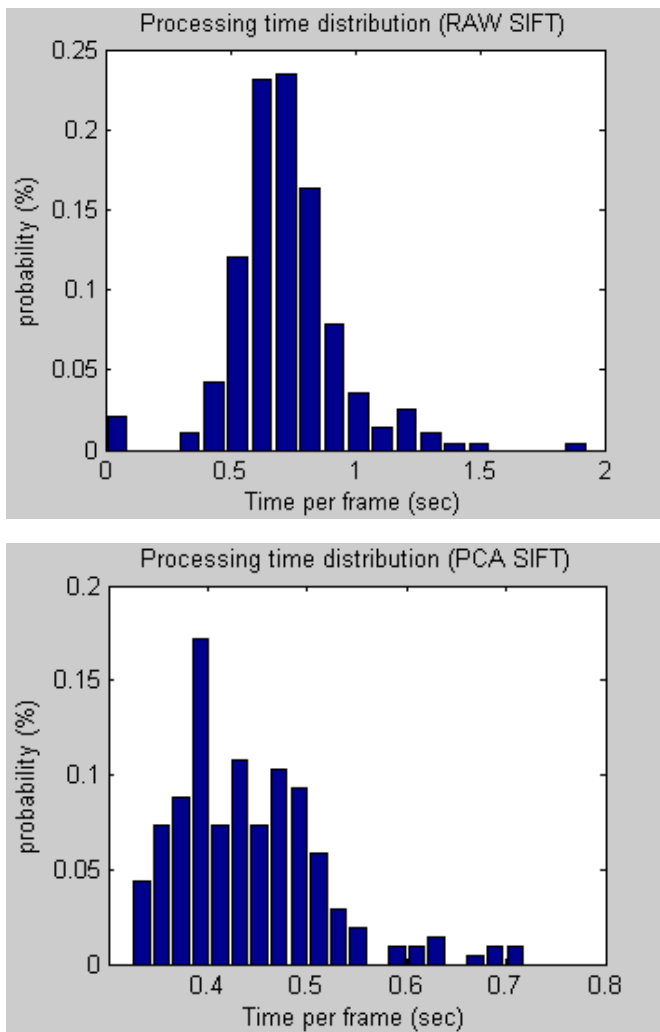


Fig. 5. Timing Analysis (20 images in database): Probability Distribution of execution time for each frame using (top figure) (a) RAW 128 dimensional SIFT features and (bottom figure) (b) PCA-SIFT using first 20 principal components. Results from processing about 200 frames.

results from the RAW SIFT features, and Fig. 5b illustrates PCA-SIFT results.

C. Timing Analysis: Variable Database Size and Principal Components

This timing analysis experiment was conducted to evaluate the scalability of the lost-robot algorithm with increasing database size. As expected, this approach is not computationally constant as map size increases. The goal is to keep execution time as close to real time as possible, where at least 1 Hz processing is desired, which would allow reasonable motion to the robot while operating in lost mode. This experiment also evaluates timing performance with various PCA settings, as we will have to balance the tradeoff between timing performance and adequate classification rates for false positives and false negatives.

As Fig. 6 illustrates, performance is linear with respect to the database size, and the PCA representation of the SIFT landmarks is able to maintain real time performance of at least

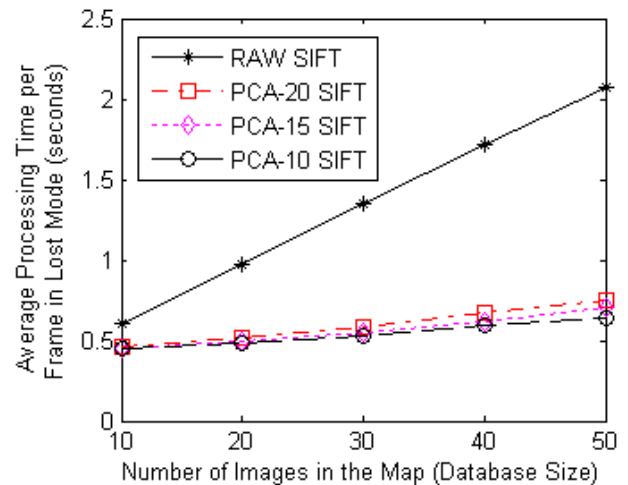


Fig. 6. Timing Analysis: Variable Database Size and Variable PCA setting for dimensionality reduction.

1 Hz with a map containing up to 50 images in the database, equivalent of a database of 4445 SIFT landmarks in memory. However, the RAW SIFT representation performs much slower as expected, as searching through 128 dimensional vectors requires significant more processing power.

D. Classification Performance

Classification performance is analyzed through testing, by evaluating the rate of false-positives and the rate of true-positives. Our goal is to minimize false-positives and maximize true-positives, which is achieved by balancing the trade-off between the two. This trade off is balanced through adjusting thresholds in the similarity function as comparison between SIFT features is accomplished through a nearest neighbors approach, where a successful match is established if a match is within a given distance from the test pattern.

Our Test data resulted in zero (0) false-positives throughout the experimentations, which consisted of testing 50 different scenes not in the map (but in the same environment) for each

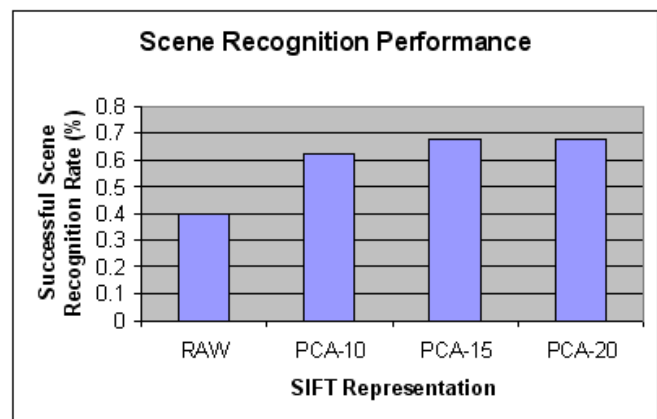


Fig. 7. Classification Performance: Successful Scene Recognition Rate for each SIFT feature representation technique.

SIFT feature representation including RAW, PCA-10, PCA-15, and PCA-20.

However, when testing for true positives, with 40 images containing scenes that are present in the map, not all images were correctly matched and identified. As Fig. 7 illustrates, 16 out of the 40 images were recognized using the RAW SIFT representation, yielding a successful scene recognition rate of only 40%. Alternately, the PCA representation of the SIFT features performed significantly better with successful recognition rates of up to 67% when using SIFT features projected onto the top 15 EigenVectors. This is a significant finding as the PCA representation of the SIFT features yield better recognition rates and better timing performance simultaneously.

VI. MONOSLAM INTEGRATION

As we have previously discussed, our lost robot localization approach included operation in dual modes, and as we have focused on the lost robot mode and demonstrated a practical implementation, further work is required to fully operate with this robust localization system. The first issue is the incorporation of SIFT transformation during MonoSLAM normal tracking operation. This change can be easily accomplished while preserving run-time performance of MonoSLAM as this would only be required while adding landmarks to the map. The landmark selection process can remain unchanged as selected by Shi and Tomasi's method [9], where only the patch/subframe from the landmark's template needs to be processed to extract the single SIFT feature descriptor. The remaining issue to incorporate our lost robot functionality into the MonoSLAM framework involves the accurate localization once a matching scene is identified, as a successful scene match does not directly represent our exact position. Further geometrical calculations from computer vision theory are required to triangulate our position and attitude if we know various points in the physical 3d space. Once we have established our state vector, we can pass the state parameters to the nominal MonoSLAM architecture, where no further calibration would be required as we already have known landmarks in the Field of View.

VII. CONCLUSION

Throughout this work, we encountered similarities between our goals and tasks commonly performed by humans on a daily basis such as navigating while driving, which is an example of the connection between human cognitive ability and robotic localization systems that involve multi-modal sensing capabilities and learning algorithms that interpret and process data in a highly efficient manner. Therefore, a closer investigation of this area may prove useful when implementing autonomous robotics applications.

Our results provide optimism and encouraging motivation as we have demonstrated a feasible implementation of lost robot localization through scene recognition using SIFT

features represented through principal components for dimensionality reduction, while performing successfully within a soft real time constraint of at least one cycle per second.

Potential expansion of this work includes the analysis for careful consideration for MonoSLAM integration, as well as additional modifications for threshold optimization to increase localization performance by increasing the scene recognition rates.

ACKNOWLEDGMENT

A special thanks to Dr. Dezhen Song for his critical role as instructor, mentor, advisor, and contributor to this study. Furthermore, we are also extremely grateful for the contributions and support by Dr. Daniele Mortari, Dr. Igor Carron, Dr. Thomas Talley, and Mr. Charles Hill.

REFERENCES

- [1] R. Smith, M. Self, and P. Cheeseman, "A Stochastic Map for Uncertain Spatial Relationships," *Proc. Fourth Int'l Symp. Robotics Research*, 1987.
- [2] J.J. Leonard, "Directed Sonar Sensing for Mobile Robot Navigation," PhD dissertation, Univ. of Oxford, 1990.
- [3] J. Manyika, "An Information-Theoretic Approach to Data Fusion and Sensor Management," PhD dissertation, Univ. of Oxford, 1993.
- [4] M. Csorba, "Simultaneous Localisation and Mapping," PhD dissertation, Univ. of Oxford, 1997.
- [5] S. Thrun and M. Montemerlo. "The GraphSLAM algorithm with applications to large-scale mapping of urban structures." *International Journal on Robotics Research*, 25(5/6):403-430, 2005.
- [6] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. "FastSLAM: A factored solution to the simultaneous localization and mapping problem." *In Proceedings of the AAAI National Conference on Artificial Intelligence*, Edmonton, Canada, 2002. AAAI.
- [7] A. Davison, "Real-Time Simultaneous Localisation and Mapping with a Single Camera," *ICCV*, 2003.
- [8] A. Davison, I. Reid, N. Molton and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM," *IEEE Trans. PAMI* 2007
- [9] J. Shi and C. Tomasi, "Good Features to Track," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 593-600, 1994.
- [10] L. Clemente, A. Davison, I. Reid, J. Neira and J. Tardós, "Mapping Large Loops with a Single Hand-Held Camera," *RSS* 2007
- [11] J. Civera, A. Davison and J. Montiel, "Inverse Depth to Depth Conversion for Monocular SLAM," *ICRA* 2007
- [12] P. Smith, I. Reid, and A. Davison, "Real-Time Monocular SLAM with Straight Lines," *BMVC* 2006.
- [13] Stephen Se, David G. Lowe and James J. Little, "Vision-based global localization and mapping for mobile robots," *IEEE Transactions on Robotics*, 21, 3 (2005), pp. 364-375.
- [14] A. Torralba, P. Sinha, "Indoor scene recognition," *AI Memo* 2001-015, *CBCL Memo* 202, 2001
- [15] David G. Lowe, "Object recognition from local scale-invariant features," *International Conference on Computer Vision*, Corfu, Greece (September 1999), pp. 1150-1157
- [16] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60, 2 (2004), pp. 91-110.
- [17] David G. Lowe, "Local feature view clustering for 3D object recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii (December 2001), pp. 682-688.
- [18] A. Oliva, A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, Vol. 42(3): 145-175, 2001.

- [19] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," *Computer Vision and Pattern Recognition*, 2004.
- [20] M. Bowling, D. Wilkinson, A. Ghodsi, "Subjective Mapping," *American Association for Artificial Intelligence*, 2006.
- [21] M. Bowling, A. Ghodsi, and D. Wilkinson, "Action Respecting Embedding," In proceedings of The 22nd *International Conference on Machine Learning (ICML 2005)*

Pedro Davalos obtained his bachelor's degree in computer engineering (2002) and is currently pursuing a masters degree in computer science, both at Texas A&M University, College Station, TX 77843, USA.

Mr. Davalos has worked on various projects as a Research Engineer at the Spacecraft Technology Center since 2002 (College Station, TX). Previous positions also include Senior Video Network Specialist at the Educational Broadcast Services Department of Texas A&M. Where his research interests include signal processing, machine learning, robotics, and computer vision. Mr. Davalos is also a member of the American Institute of Aeronautics and Astronautics.