Proceedings of 2012 IEEE
International Conference on Mechatronics and Automation
August 5 - 8, Chengdu, China

Performance Evaluation of a Pose Estimation Method based on the SwissRanger SR4000

Soonhac Hong and Cang Ye
Department of Systems Engineering
University of Arkansas Little Rock, AR 72204, USA
{ sxhong1 & cxye}@ualr.edu

Michael Bruch and Ryan Halterman
Space and Naval Warfare Systems Center Pacific
San Diego, CA 92152, USA
{michael.bruch & ryhalt}@navy.mil

***Abstract*—This paper presents an experimental study on the performance of a Pose Estimation (PE) method based on a 3D time-of-flight camera—the SwissRanger SR4000. The PE method tracks the visual features in the camera's intensity image and computes the camera's pose change from the 3D data of the matched features. To attain a small PE error, the noises of the sensor's intensity and range data are analyzed and a Gaussian filter is applied to reduce the noises. The statistical property of the filtered data is then characterized and the result is used to determine the minimum number of 3D data points that are required for a satisfactory PE accuracy. Two feature extractors, the SIFT (Scale Invariant Feature Transform) and SURF (Speed Up Robust Features) extractors, are used for the PE method and their performances are compared in term of PE error and computational time.**

Experimental results with various combinations of rotation and translation movements demonstrate that the SIFT extractor outperforms the SURF extractor in both PE accuracy and repeatability.

Index Terms* - *Pose Estimation, Time-of-Flight Camera, Feature Descriptor, Visual Odometry.

I. INTRODUCTION

Pose Estimation (PE) is an essential capability for a mobile robot to perform its critical functions including localization and mapping. In a 3D space, a robot pose refers to its attitude (roll, pitch, yaw angles) and position (*X*, *Y*, *Z* coordinates). This is referred to as 6D pose. In an environment without the support of navigational infrastructure (GPS, active beacons, etc.), a robot may calculate its pose via dead reckoning by using data from an Inertial Measurement Unit (IMU) and wheel odometry or by tracking the features detected by its vision sensor in the operating environment. The former is a proprioceptive approach as it does not use any external reference and the latter is an exteroceptive one as it tracks environmental features (external references) over time. The measurement accuracy of an IMU is subject to bias drifts of its motion sensors (accelerometers) and rotation sensors (gyros) that accrue pose errors in over time. Wheel odometry alone cannot provide a reliable estimate of movement due to wheel slip. A PE system based on an IMU and/or wheel odometry may accumulate a large pose error with time and may completely fail in case of excessive wheel slip (e.g., when a robot moves on loose soil).

Exteroceptive PE methods are less prone to these problems as they use static external references to determine changes in robot pose. LADARs [1]-[3] have been widely used for PE. A 3D LADAR [3] is currently costly and is not suitable for a small robot due to its large dimension. When a 2D LADAR is used for 6D pose estimation, a scanning mechanism is needed to rotate the entire sensor for generating 3D data [1], [2]. Such a system has a low frame rate and is often too bulky for a small robot. Most of the existing LADAR-based PE methods employ the Iterative Closest Point (ICP) or its variants [1]-[3]. The ICP approach requires an initial guess of pose that moves the data close enough to the model to alleviate the local minimum problem. It cannot find a solution when the environment is featureless (e.g., a flat surface).

Visual features are often more available in the operating environment of a robot. They may even exist in a geometrically featureless environment. The representative approach to PE with visual features is Visual Odometry (VO) [4]-[9], which employs a stereovision system to estimate ego-motion by detecting and tracking the visual features in the stereo image pairs from frame to frame. The 3D positions of the features in each frame are determined by stereo matching. Feature tracking is performed by spatial correlation search. The search methods [4], [5] require an initial estimate of the motion that is obtained from wheel odometry and inertial sensors. Feature matching can also been achieved by using feature descriptors [6], [7]. This feature matching approach does not perform a spatial correlation search and thus does not require an initial motion estimate. However the feature descriptors need to be salient and robust to scale change and rotation in order for the VO method to work well with a large frame-to-frame movement. Although the VO method is suitable for a small robot, the use of a stereovision system imposes significant limitations on autonomous robot navigation. First, the depth measurement error of a stereovision system grows quadratically with the true distance, meaning that the VO's PE accuracy drop off quickly with increasing distances of feature points. Second, stereovision systems cannot produce a fully dense set of depth data of the scene due to featureless image regions. As a result, stereovision data is not reliable for obstacle avoidance. This problem is usually resolved by using additional perception sensor(s) such as a LADAR. This multi-sensor approach is, however, not suitable for navigating a small robot.

Commercially available Flash LIDAR now has sufficient accuracy for robotic application. A Flash LIDAR simultaneously produces intensity and range images of the scene at a video frame rate. It has the following advantages over stereovision systems: (1) it measures depth by Time-Of-Flight (TOF) and therefore has consistent measurement accuracy in its full range; and (2) it may return

978-1-4673-1278-3/12/$31.00 ©2012 IEEE

fully dense depth data across its field-of-view. The commercially available Flash LIDAR includes the SwissRanger [17] and TigerEye 3D [18] cameras. In [11], [12], a sensor suite that pairs a SwissRanger 3D camera with a conventional video camera was used for PE. SIFT features were extracted from the video image. The corresponding 3D points of the SwissRanger's range data were determined by an interpolation and search process. The data association process incurs computation time and a precise pixel-to-pixel match cannot be achieved. To overcome these drawbacks, the authors of this paper introduced a PE method [14] based on a single 3D camera—the SwissRanger SR4000. In the PE method, SIFT features were extracted from the SR4000's intensity image and the features' corresponding range data were used as the 3D data points for PE. This is a precise pixel-to-pixel match and no search process is needed. Since the method simultaneously uses Visual and Range data for PE, it was termed VR-Odometry (VRO). The work in [14] only presents a proof of concept study of the VRO that uses the sensor's raw data. A similar PE method was presented in [13] where the SURF descriptor was employed for feature detection and matching. The motivation was that the SURF descriptor has similar performance to SIFT but is much faster according to the recent study [15]. However, it is not clear from [13] if the PE accuracy and computational cost of the SURF-based method are satisfactory. In [11], [13] the SwissRanger SR-2 and SR-3000 were characterized. The uncertainty of each pixel's range measurement was related to its intensity value and the bias of the range measurement was related to the intensity and true range values. These statistical data were then used by an information filter to estimate the robot pose over multiple time steps.

In this paper we investigate: (1) the impacts of the SR4000's noises to the VRO's PE error; and (2) the performances of the SIFT and SURF descriptors in terms of PE error and computational cost. The objective is to allow the VRO to provide sufficiently accurate PE for a portable blind navigational device [14] that uses a SR4000 instead of a Kinect sensor due to the SR4000's smaller footprint.

This paper is organized as follows. Section II briefly introduces the SwissRanger SR4000. Section III gives an overview on the VRO method. Section IV describes the experimental setup for the benchmark study. Section V characterizes the noises of the sensor's data. Section VI presents a simulation approach for determining the minimum number of matched points required for the VRO. Section VII evaluates the performance of the SIFT-based VRO with individual movements. Section VIII compares the SIFT and SURF extractors' performances in the context of PE. Section IX further evaluates the SIFT-based VRO's PE performance. The paper is concluded in Section X.

II. SwissRanger SR4000

The SwissRanger SR4000 (Fig. 1(b)) is a 3D TOF camera. The SR4000 operates like a 2D digital camera, but measures the range to all objects in its field of view (the scene). The camera illuminates the scene with modulated infrared light and focuses the image onto the 3D camera's Focal Plane Array (FPA). Each pixel on the FPA measures the TOF, and thus the distance, based on phase shift. The final result is a fully dense range image. The camera also produces an intensity image simultaneously. The camera is small in size ($65\times65\times68$ mm^3). But it is able to produce dense 3D data points (176×144 =25,344 points) per frame at a frame rate up to 50 Hz. It has a 43˚ (horizontal) by 34˚ (vertical) field of view and a 0.24˚ angular resolution.

III. VR-Odometry

The VRO employs a local feature detector to extract features in current intensity image and match them to the features in the next intensity image by the feature descriptors. As the features' 3D coordinates are known, the feature-matching process results in two 3D data sets, $\{p_i\}$ and $\{q_i\}$ for $i=1,\cdots,N$. The rotation and translation matrices, R and T, between these two image frames can be determined by minimizing the error residual $e^2=\sum_{i=1}^{i=N}\left\|p_i-Rq_i-T\right\|^2$. This least-squares data fitting problem can be solved by the Singular Value Decomposition (SVD) method as described in [16]. As the feature-matching based on the local descriptors may result in incorrect correspondences (outliers), a RANSAC (Random Sample Consensus) process is implemented to reject the outliers. The entire method is as follows:

1) Detect features in two consecutive intensity images, find the matched features based on the feature descriptors and form the corresponding 3D data sets $\{p_i\}$ and $\{q_i\}$. Repeat steps 2 & 3 for k, $k=1,\cdots,K$, times until a termination criteria is met.

2) Draw a sample by randomly selecting 4 associated point-pairs from the two data sets to form $\{p_m\}$ and $\{q_m\}$ for $m=1,\cdots,4$. Then find the least-squares rotation and translation matrices (R_k and T_k) for $\{p_m\}$ and $\{q_m\}$.

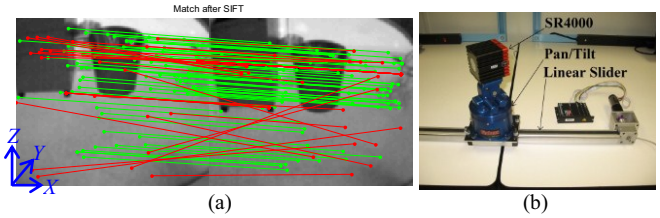
3) Project the entire data set $\{q_i\}$ onto $\{p_i\}$ by the found transformation and compute error $e_i^2=\left\|p_i-R_kq_i-T_k\right\|^2$; $i=1,\cdots,N$ for each point-pair. A threshold τ is used to score the support S_k for this transformation (R_k and T_k): S_k is incremented once for each $e_i^2<\tau$.

4) The transformation with the highest score is recorded. The corresponding data sets $\{p_j\}$ and $\{q_j\}$ for $j=1,\cdots,S_k$ (called inliers, of which each data-pair past the threshold test of step 3) are selected and used to compute the maximum likelihood estimate of $\hat{R}$ and $\hat{T}$ of the camera by the SVD method. The camera's Euler angles are computed from $\hat{R}$ and its translation is determined by $\hat{T}$.

Given the true inlier ratio ε, the minimum number of repetitions required to ensure, with some level of confidence η, that data sets $\{p_j\}$ and $\{q_j\}$ are outlier-free, can be computed by

$$K_{min} = \frac{log(1-\eta)}{log(1-\varepsilon^m)}. \quad (1)$$

This number is then used as general termination criteria [19] to terminate the RANSAC process. In this paper m=4 and η=0.99 are used. In fact, the true inlier ratio ε is a priori unknown. An estimate on this ratio can be found by using the sample which currently has the largest support. This estimate is then updated as the RANSAC process proceeds. The total computational time of the RANSAC process is determined by the values of K_{min} and N (the total number of the 3D data points). Figure 1(a) depicts the matched features that are classified into inlier (green) and outlier (red) by the RANSAC process.



(a) (b)

Fig. 1 (a) RANSAC process for outlier removal: Features matched by SIFT descriptor consists of inliers (in green) and outliers (in red) which is identified and removed by the RANSAC process. (b) Experimental setup: the SwissRanger SR4000 mounted on the motion table

IV. Experimental Setup

An in-house built motion table (Fig. 1(b)) is used to produce ground truth rotation and translation. The motion table consists of a TRACLabs Biclops pan/tilt unit and a Techno Linear Slider driven by a servo motor. The motion table has a ±60° tilt and ±180° pan ranges with an angular resolution of 0.018°. It has a 3-meter linear movement range with ~0.01 mm resolution.

A desktop computer with a 3.30GHz AMD Phenom™ II X6 1000T processor and 4 GB memory is used to process the intensity and range data. All statistical results are computed using 540 data frames. This sample number is determined based on the Central Limit Theorem and the statistical results of the VRO in our previous paper [10].

V. Noise Analysis

In our previous work [10], we have demonstrated that the PE error of the VRO using raw sensor data follows a Gaussian distribution. Noises in the SR4000's range data directly contribute to the VRO error while noises in the intensity data may result in error in feature detection and eventually produce error in PE. To analyze the noises of the SR4000's intensity and range data, we take 540 image frames from the sensor in a typical office environment and plot the distributions of the intensity and range values of each pixel. It is found that both intensity and range values follow Gaussian distributions, meaning that the noises in the intensity and range data are Gaussian noises.

A 3×3 low-pass Gaussian filter (σ=1) is applied to the intensity and range images to reduce the noises. The distribution plots of the data before and after applying the Gaussian filter demonstrates that the filter substantially reduces the standard deviations with slight changes in the mean values. To quantify the noise levels of the raw and filtered data, we compute the noise ratio (i.e., the ratio of the standard deviation to the mean) of each pixel in the intensity and range data. Table I shows the maximum, mean and minimum noise ratios in a typical data frame. It can be seen that the Gaussian filter reduces the overall noise levels (the mean noise ratio) of the intensity and range data by 57% and 63%, respectively. We have carried out experiments with different scenes and the results are similar.

TABLE I Noise Ratio of Raw and Filtered SR4000 Data

Noise Ratio (%)		Raw Data	Filtered Data	Noise Reduction
Intensity	Maximum	20.66	9.88	52%
	Mean	2.28	0.99	57%
	Minimum	0.47	0.42	11%
Range	Maximum	14.66	7.96	46%
	Mean	1.50	0.55	63%
	Minimum	0.31	0.12	61%

The maximum, mean and minimum is computed from 25344 (176×144) pixels.

VI. Minimum Number of Matched Points for VRO

In theory, the pose change between two image frames can be computed from three matched data points. This scheme only works well with noise-free range data. When the range data contain noise, more matched data points are required to attain a sufficiently small PE error. In this work, this issue is investigated through simulation experiments.

Let a noise-free point set be denoted by $\{p_i\}$; $i = 1,\cdots,N$. A predetermined transformation, i.e., a pose change given by (θ, ϕ, ψ, x, y, z), is applied to $\{p_i\}$ to generate a new data set $\{q_i\}$. This means that data points in $\{p_i\}$ are matched to those in $\{q_i\}$ by the given rotation and translation. Gaussian noise is then added to each point in $\{p_i\}$ and $\{q_i\}$. This produces data sets $\{p_i'\}$ and $\{q_i'\}$ that simulate the SR4000's range data. The transformation matrices, R' and T', between $\{p_i'\}$ and $\{q_i'\}$ can be computed by the SVD method described in Section III. The corresponding pose change (ϕ', θ', ψ', x', y', z') are then determined and treated as the pose measurement. To reflect the noise characteristics of the SR4000's range data, we used the sensor's maximum, mean and minimum noise ratios (in Table I) to generate the simulated range data. We then computed the pose measurement error δ=(δ_ϕ, δ_θ, δ_ψ, δ_x, δ_y, δ_z,)=(ϕ'-ϕ, θ'-θ, ψ'-ψ, x'-x, y'-y, z'-z) for each case using different number of matched points. Each element of δ and the absolute value of the derivative $|d\delta|$ are plotted against the number of matched points that are used for computing the pose change.

The results show that: (1) Each element of δ decreases gradually with increasing number of matched points; and (2) each element of $|d\delta|$ decreases with the number of matched points. More importantly, $|d\delta_\phi|$, $|d\delta_\theta|$ and $|d\delta_\psi|$ stay within the SR4000's angular resolution (0.24°) and $|d\delta_x|$, $|d\delta_y|$ and $|d\delta_z|$ are within the SR4000's absolute accuracy (10 mm) when the number of matched points is not smaller than 12. Therefore, 12 are selected as the Minimum Number of Matched Points (MNMP) for PE. Figure 2 shows the pitch derivative plots. As

the similar results are shown in other movements, the other plots are omitted for conciseness.

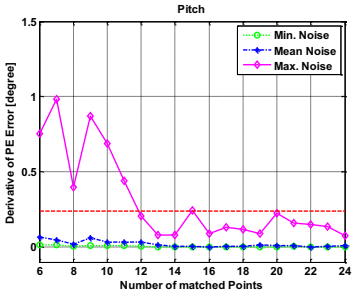


Fig. 2 Derivative of PE error $|d\delta_\theta|$ versus number of matched points under different noise levels. The solid red line is the SR4000's angular resolution.

It is noted that the MNMP depends on the threshold values. A larger/smaller threshold value results in a smaller/larger MNMP. Decreasing the MNMP will increase the success rate of PE with a larger PE error. Therefore, the MNMP should be carefully selected according to the system specification. The MNMP may be used to indicate a failure of the VRO when the number of initial matched featured (based on the feature discriptors) is smaller than the MNMP. In this case, the VRO should skip the current frame of data and move to the next frame, or be replaced by an auxilliary PE means, if available.

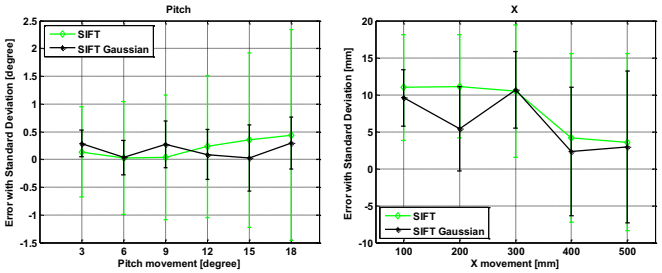
VII. PE Error and Computational Time

A. Accuracy and repeatability of VRO

Since the Gaussian filter drastically reduces the noise of the SR4000's data, it is anticipated that the filtering process substantially reduces the VRO's PE error. Experiments are performed to examine the PE error (with/without the Gaussian filter) for each individual motion. The SIFT feature extractor is used in the experiments. In the first experiment, the SR4000 undergoes pitch movements over the range [3°, 18°] (step size: 3°). 540 frames are captured before and after each pitch rotation for computing the pose change. Figure 3(a) shows the error bar plot of the pitch measurements.

It can be observed that the use of the Gaussian filter resulted in a much better repeatability of pitch estimation: the standard deviation was reduced by 1.7~2.9 times. The Gaussian filter slightly shifts the mean error. This change is within [-0.3°, 0.3°]. It is small compared with the reduction in the standard deviation, which is 1.3°~2.9°. We carry out similar experiments to test the VRO errors with roll, yaw, X and Y movements. The result of X movement is depicted in Fig. 3(b).

For roll measurement, the Gaussian filter slightly reduced the mean errors but retained similar standard deviations. Both the mean errors and standard deviations are within the SR4000's angular resolution (0.24°). For yaw measurement, the Gaussian filter causes significant reductions in the standard deviations but slightly shifts the mean errors, a similar result as observed in the pitch measurement. Again, both the mean errors and standard deviations are within the SR4000's angular resolution. One common feature in the pitch and yaw measurements is that the standard deviations tend to grow with increasing rotation angle.



(a) (b)

Fig. 3 Error bars of pitch and X measurements. The mean error of each measurement is plotted with an error bar of ±1σ. σ is the standard deviation.

For X measurement, both the mean errors and standard deviations are reduced by the Gaussian filter. Their values are around the absolute accuracy of the SR4000 (±10 mm). The standard deviation of the VRO with Gaussian filter grows with increasing X movement. For Y measurement, the Gaussian filter lowers the standard deviations but slightly increases the mean errors. However, both of them are within the absolute accuracy of the SR4000. It can be observed that the measurement errors of X movements are larger than that of Y movements. This is because the same amount of movement in X axis results in a larger movement of features in the image plane (equivalent to a larger pitch or yaw movement) than it does in Y axis and hence a large measurement error.

The quantitative results of the measurement accuracy and repeatability of the VRO with the Gaussian filter are tabulated in Tables II and III. It is noted that for the 2nd group of data in Table II (i.e., the sensor undergoes pitch movements), some of the pitch/yaw measurements incur larger mean errors and

TABLE II PE Errors in Rotation Measurement

MV: (μ, σ) / TV: (ϕ, θ, ψ)	Roll ϕ (°)	Pitch θ (°)	Yaw ψ (°)
(3, 0, 0)	(0.17, 0.11)	(0.06, 0.07)	(0.04, 0.06)
(6, 0, 0)	(0.16, 0.10)	(0.02, 0.06)	(0.03, 0.07)
(9, 0, 0)	(0.07, 0.10)	(0.07, 0.06)	(0.05, 0.06)
(12, 0, 0)	(0.02, 0.11)	(0.09, 0.07)	(0.01, 0.07)
(15, 0, 0)	(0.00, 0.10)	(0.05, 0.08)	(0.11, 0.09)
(18, 0, 0)	(0.01, 0.11)	(0.12, 0.08)	(0.03, 0.07)
(0, 3, 0)	(0.07, 0.11)	(0.30, 0.25)	(0.05, 0.25)
(0, 6, 0)	(0.00, 0.13)	(0.03, 0.31)	(0.46, 0.33)
(0, 9, 0)	(0.07, 0.15)	(0.27, 0.41)	(0.56, 0.35)
(0, 12, 0)	(0.04, 0.17)	(0.09, 0.46)	(0.81, 0.42)
(0, 15 0)	(0.16, 0.21)	(0.02, 0.56)	(0.95, 0.49)
(0, 18, 0)	(0.14, 0.23)	(0.29, 0.49)	(0.61, 0.46)
(0, 0, 3)	(0.02, 0.07)	(0.09, 0.13)	(0.17, 0.11)
(0, 0, 6)	(0.02, 0.08)	(0.09, 0.14)	(0.21, 0.11)
(0, 0, 9)	(0.01, 0.08)	(0.18, 0.16)	(0.14, 0.16)
(0, 0, 12)	(0.03, 0.09)	(0.18, 0.2)	(0.15, 0.22)
(0, 0, 15)	(0.01, 0.12)	(0.22, 0.22)	(0.23, 0.27)
(0, 0, 18)	(0.01, 0.11)	(0.2, 0.21)	(0.17, 0.28)

MV: Measured Values, TV: True Values, μ: mean error, σ: standard deviation.

TABLE III PE Errors in Translation Measurement

MV: (μ, σ) / TV: (X, Y, Z)	X (mm)	Y (mm)	Z (mm)
(100, 0, 0)	(9.8, 4.0)	(0.4, 1.4)	(3.3, 2.4)
(200, 0, 0)	(5.6, 5.5)	(2.7, 1.7)	(3.9, 2.9)
(300, 0, 0)	(10.5, 5.2)	(3.4, 1.6)	(7.7, 3.6)
(400, 0, 0)	(2.8, 8.9)	(4.7, 2.7)	(6.9, 6.8)
(500, 0, 0)	(3.4, 10.7)	(5.5, 2.5)	(0.8, 7.3)
(0, 100, 0)	(1.4, 2.8)	(4.3, 1.7)	(3.4, 2.7)
(0, 200, 0)	(2.8, 2.8)	(6.2, 1.7)	(2.5, 3.1)
(0, 300, 0)	(0.9, 2.7)	(7.7, 1.8)	(0.3, 3.5)
(0, 400, 0)	(3.3, 3.1)	(9.5, 1.8)	(3.3, 3.7)
(0, 500, 0)	(12.6, 4.8)	(8.5, 2.5)	(6.1, 4.9)

MV: Measured Values, TV: True Values, μ: mean error, σ: standard deviation

standard deviations. We find that this is because the data has larger noise ratios. We will further investigate this issue and attempt to improve the results in our future work.

B. Computational time of VRO

We find in our experiments that the Gaussian filter increases the number of SIFT features in each filtered intensity image and resulted in a much larger number of initially matched features with a higher inlier ratio. Therefore, the computational time for feature extraction increases. The RANSAC computational time decreased due to higher inlier ratio. However, it accounts for a much smaller portion of the total VRO computational time. As a result, the use of the Gaussian filter results in a much higher computational time for the VRO. Figure 4 shows the numbers of inliers and outliers and computational times of the feature extraction and RANSAC processes of the VRO with/without the Gaussian filter in the cases with various pitch movements.

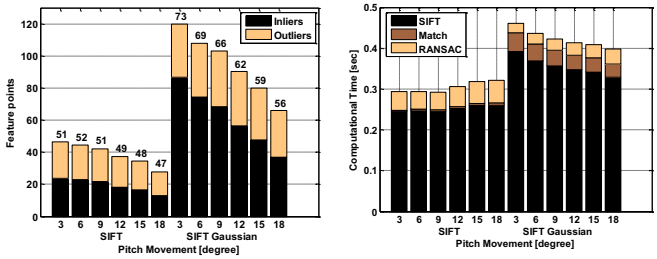


Fig. 4 Inlier and outlier numbers and the computational time of the SIFT-based VRO with/without the Gaussian filter

VIII. Performances of Feature Extractors

The SURF is a representative robust feature descriptor with a much better computational efficiency than the SIFT. It has been used for robot PE in the literature [13]. However, no quantitative study on its PE performance has been reported so far. In this section, we compare the performance of the SURF-based VRO with that of the SIFT-based VRO. In both cases, the Gaussian filter was used to reduce the noise ratios of the data.

A. Accuracy and repeatability

We compare the two VROs' mean errors and standard deviations through extensive experiments with individual movements (roll, pitch, yaw, *X* and *Y*). Figure 5 shows two of the experimental results. It can be observed that the SIFT-based VRO exhibits a much better accuracy and repeatability in PE. The only exception is the *Y* measurement where the SURF-based VRO has slightly smaller mean errors

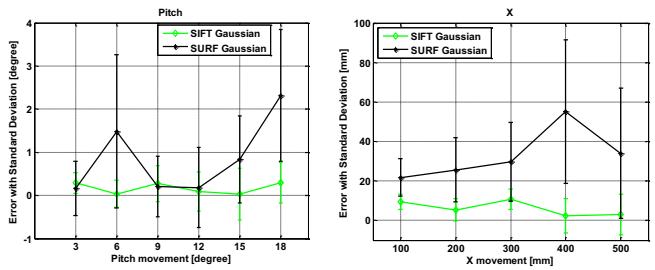


Fig. 5 Comparison of the SIFT-based and SURF-based VROs' errors in pitch and *X* movement

and the SIFT-based VRO has smaller standard deviation. However, both methods' PE errors are within the SR4000 absolute accuracy. This means that they have similar performance in measuring *Y* movement.

A closer examination into the SURF-based VRO reveals that: (1) For the measurements in *Y* movements, the SURF feature extractor produces sufficient number of inliers and thus results in PE errors comparable to that of the SIFT-based VRO; (2) In all other cases, the number of inliers produced by the SURF feature extractor is lower than the MNMP and therefore, the use of the SURF descriptors results in much larger PE errors.

B. Computational time

Figure 6 depicts the inliers and outlier numbers and the run time of the SIFT-based and SURF-based VROs under the conditions with pitch movements in range [3°, 18°]. It can be seen that the SIFT extractor produced many more features with a much higher inlier ratio than the SURF counterpart for each pitch movement. Also, the computational time of the SURF is ~200 ms smaller than that of the SIFT. For the SURF-based VRO, the RANSAC computational time may drastically increase and outnumber the computational time of feature extraction when the inlier ratio is very low (e.g., the case with 18° pitch movement). We obtain similar results in the cases with other individual movements.

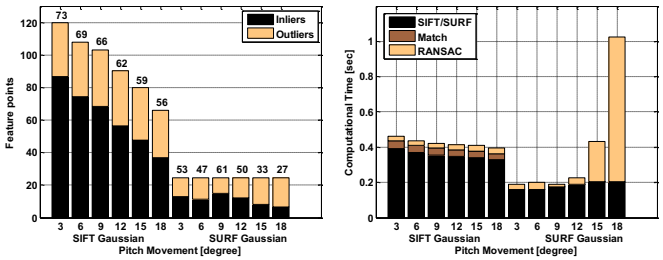


Fig. 6 Inlier and outlier numbers (top) and the computational time (bottom) of the SIFT-based and SURF-based VROs with pitch movements. The numbers on top of the bar graph in the left are inlier ratios in percentage.

In general, the use of the SURF may save some computational time in the feature extraction for the VRO at an unacceptable cost of losing too much PE accuracy and repeatability.

The thresholds of the SIFT and SURF detectors used in the experiments are 0.0067 and 0.0001, respectively. By decreasing the Hessian response threshold of SURF, we got a larger number of SURF features. But the inlier number did not increase proportionally. This resulted in a lower inlier ratio and thus an increased RANSAC computation time. The increase of feature number did not improve pose estimation performance. We found similar phenomenon with SIFT detector. This indicates that a feature detector's robustness to scale and rotation is the determining factor of its pose estimation performance.

IX. PE Error with Combinatory Motion

Since the SIFT feature extractor outperforms the SURF extractor, the SIFT-based VRO was selected for PE. Its PE

accuracy and repeatability are further investigated under conditions with combinatory motion (both rotation and translation). The results are tabulated in tables IV & V.

The first combinatory motion consists of pitch, yaw and *Y* movements. Like the results with individual movements (in table II), most of the mean errors and standard deviations of the SIFT-based VRO are within the SR4000's angular resolution and absolute accuracy.

The second combinatory motion consists of pitch, yaw and *X* movements. While most of the standard deviations are within the sensor's angular resolution and absolute accuracy, many of the mean errors go beyond the boundaries. This conforms to our earlier finding that an *X* movement results in a larger PE error than a *Y* movement does. However, the overall PE accuracy and repeatability are still good.

Since a robot's forward movement (along *Y* axis) usually dominates its side movement (along *X* axis), we believe that the measurement performance in Table IV has more significant meaning to robot pose estimation.

TABLE IV MEASUREMENT OF ROTATION (θ, ψ) AND TRANSLATION (Y)

MV: (μ, σ) / *TV*: (ψ, θ, Y)	Pitch θ (°)	Yaw ψ (°)	Y (mm)
(-3, 9, 390)	(0.45, 0.07)	(0.12, 0.06)	(1.23, 1.12)
(-6, 3, 196)	(0.31, 0.07)	(0.02, 0.07)	(3.52, 1.49)
(-9, 12, 386)	(0.15, 0.22)	(0.16, 0.20)	(3.61, 1.93)
(-12, 6, 593)	(0.28, 0.41)	(0.26, 0.57)	(10.90, 4.07)

MV: Measured Values, TV: True Values, μ: mean error, σ: standard deviation.

TABLE V MEASUREMENT OF ROTATION (θ, ψ) AND TRANSLATION (X)

MV: (μ, σ) / *TV*: (ψ, θ, X)	Pitch θ (°)	Yaw ψ (°)	X (mm)
(3, 12, 99)	(0.05, 0.08)	(0.55, 0.13)	(13.82, 5.05)
(6, 9, 198)	(0.28, 0.10)	(0.78, 0.21)	(19.41, 6.91)
(9, 6, 287)	(0.16, 0.20)	(0.80, 0.35)	(22.17, 11.83)
(12, 3, 395)	(0.22, 0.20)	(0.90, 0.39)	(28.75, 14.26)

MV: Measured Values, TV: True Values, μ: mean error, σ: standard deviation.

X. CONCLUSIONS

In this paper, we have evaluated the performance of the SR4000 based pose estimation method, called VR-Odometry (VRO). We first evaluated the noise ratios of the sensor's raw data, including intensity and range data, and applied a Gaussian filter to reduce the noises of the data. We then characterized the noise characteristics of the filtered data. Based on the characteristics, we determined the minimum number of matched points that allows for a sufficiently small PE error. Our experimental study has demonstrated that the Gaussian filtering substantially improves the VRO's measurement accuracy and repeatability. To select a suitable feature extractor for the VRO, we evaluated the PE performances of the SIFT and SURF descriptors and the experimental results revealed that the SIFT descriptor outperformed the SURF descriptor in PE. Our results have also demonstrated that the SIFT-based VRO's PE accuracy and repeatability mostly stay within the SR4000's angular resolution and absolute accuracy when the sensor undergoes an individual rotation/translation. This is also true for the case with a combinatory movement if the linear movement along *Y* axis prevails (the usual case in robot navigation). The VRO method is well suited for small robot and robotic device where both size and PE accuracy are determining factors.

In terms of future research directions, we will look into the possibility of reducing the computational time for feature extraction. We will also develop a method to increase the inlier ratio of the matched features by removing the majority of the outliers through a non-RANSAC procedure. This may speed up or even remove the RANSAC process.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation (IIS-1017672) and NASA (NNX09A072A).

REFERENCES

[1] D. Cole and P. Newman, "Using laser range data for 3D SLAM in outdoor environments," *Proc. IEEE International Conference on Robotics and Automation*, 2006, pp. 1556-1563.

[2] O. Wulf, A. Nüchter, J. Hertzberg, and B. Wagner, "Benchmarking urban six-degree-of-freedom simultaneous localization and mapping," *Journal of Field Robotics*, vol. 25, no. 3, pp. 148-163, 2008.

[3] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Visually bootstrapped generalized ICP," *Proc. IEEE International Conference on Robotics and Automation*, 2011, pp. 2660-2667.

[4] Yang Cheng, M. W. Maimone and L. Matthies, "Visual odometry on the Mars exploration rovers – a tool to ensure accurate driving and science imaging," *IEEE Robotics & Automation Magazine ,* vol. 13, no. 2, pp. 54-62, 2006.

[5] M. Agrawal, K. Konolige, R. Bolles, "Localization and mapping for autonomous navigation in outdoor terrains: a stereo vision approach," *Proc. IEEE Workshop on Applications of Computer Vision,* 2007.

[6] D. Nister, O. Naroditsky and J. Bergen, "Visual odometry," *Proc. IEEE Conference on Computer Vision Pattern Recognition*, 2004, pp. 652-659.

[7] H. Hirschmüller, P. Innocent, and J. Garibaldi, "Fast, unconstrained camera motion estimation from stereo without tracking and robust statistics," *Proc. International Conference on Control, Automation, Robotics and Vision*, 2002, pp. 1099-1104.

[8] A. I. Comport, E. Malis and P. Rives, "Accurate quadrifocal tracking for robust 3D visual odometry," *Proc. IEEE International Conference on Robotics and Automation*, 2007, pp. 40-45.

[9] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems,* 2008, pp. 3946-3952.

[10] C. Ye and M Bruch, "A visual odometry method based on the SwissRanger SR-4000," *Proc. Unmanned Systems Technology XII Conference at the 2010 SPIE Defense, Security, and Sensing Symposium.*

[11] L.P. Ellekilde, S. Huang, J.V. Miró, and G. Dissanayake, "Dense 3D map construction for indoor search and rescue," *Journal of Field Robotics*, vol. 24, pp. 71-89, 2007.

[12] W. Zhou, J. V. Miró, and G. Dissanayake, "Information-Driven 6D SLAM Based on Ranging Vision," *Proc. EEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 2072-2077.

[13] J. J. Wang, G. Hu, S. Huang, and G. Dissanayake, "3D Landmarks Extraction from a Range Imager Data for SLAM," *Australasian Conference on Robotics and Automation (ACRA)*, 2009.

[14] C. Ye, "Navigating a Portable Robotic Device by a 3D Imaging Sensor," *Proc. EEEE Sensors Conference,* 2010, pp. 1005-1010.

[15] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Computer Vision–ECCV*, 2006, pp. 404-417.

[16] K. S. Arun, et al, "Least square fitting of two 3-d point sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 5, pp. 698-700, 1987.

[17] http://www.mesa-imaging.ch/

[18] http://www.advancedscientificconcepts.com/products/tigereye.html

[19] R. Raguram, J. M. Frahm, and M. Pollefeys, "A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus," *Computer Vision–ECCV*, 2008, pp. 500-513.