Chapter 4

Experimental Design

To measure the performance of our proposed ST-SLAM algorithm, we have set up two different experiments: ST-SLAM in a simulated large-scale outdoor environment and ST-SLAM in a real indoor environment. Through these two experiments, we attempt to empirically demonstrate ST-SLAM system's scalability in both space and noise level. We describe the details of the experiments in the following sections.

4.1 ST-SLAM in a Simulated Large-Scale Outdoor Environment

In this experiment, we attempt to demonstrate the ST-SLAM system's adaptability to large-scale environments. We have conducted the entire experiment using simulation including the environment, a mobile robot, and the vision-based sensor system equipped on the robot.

We chose, as an outdoor testbed, a publically-available 3D model of Housestead's fort, a Roman garrison from the 3rd century A.D. on Hadrian's Wall in Britain [30]. This is a 100 m $\times$ 55 m large-scale outdoor environment with roads, buildings and walls. Besides its scale, it is a challenging testbed for ST-SLAM due to a variety of objects with different kinds of texture. The ground of the environment has a smooth gradation of color intensity. Building surfaces have a grid-like repetitive patterns in outer appearance, which adds a difficulty to establish 2D corner point correspondences across trinocular images. The walls surrounding the fort has continuous salt-and-pepper texture, which also brings a hard correspondence problem of 3D points observed at different time.

We simulated a teleoperated robot making a round trip of a 100 m $\times$ 55 m rectangular region approximately at speed 1 m/s. We show a bird-eye view of the ground truth robot path in Fig. 4.1. We denote the robot pose at time t in the environment using a six-vector $\boldsymbol{s}_t = [s_{x,t}, s_{y,t}, s_{z,t}, s_{\phi,t}, s_{\theta,t}, s_{\psi,t}]^T$, where the first three components represent 3D robot position in world-coordinates and the last three components represent the pich, roll and yaw orientations of the robot relative to the world. $\boldsymbol{s}_{0:t}$, the entire path of the robot from time 0 to time t is a set of robot poses $\{\boldsymbol{s}_i,$ where $i = 0, 1, \ldots, t\}$. The robot observes the environment using its trinocular vision sensor at each discrete time i and plans the next robot-relative move $\boldsymbol{u}_{i+1}$, which we call robot *odometry*. On the other hand, we write the resulting *true move* in robot-relative coordinates as $\Delta\boldsymbol{s}_{i+1}$. In this simulation experiment, the time interval of each move was a constant $\Delta t = 1$ for all moves in simulation.

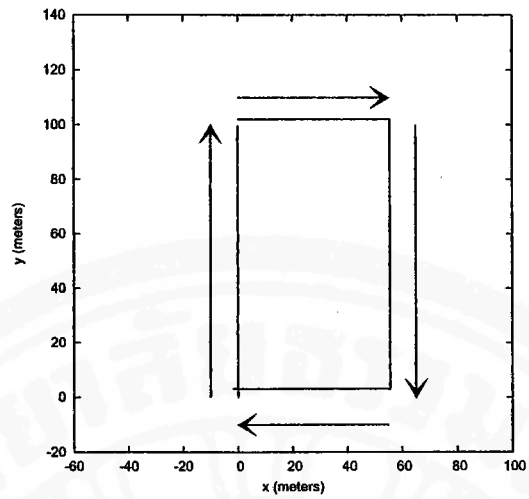


Figure 4.1: Ground truth path of the simulated robot. The path is approximately a 100 m × 55 m rectangle. The robot departs from the origin following the directions indicated by arrows and comes back to the same point.

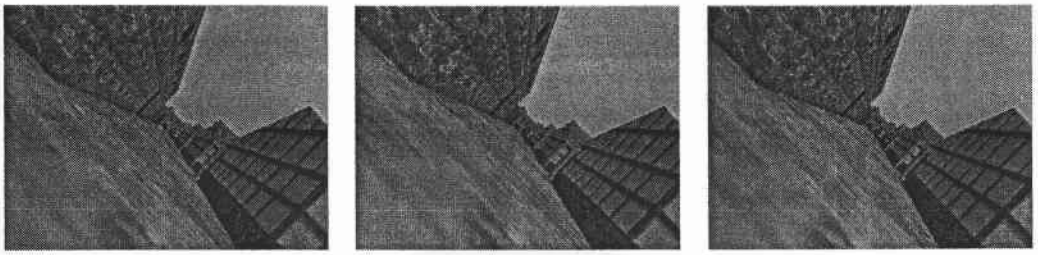
The robot's motions were either mostly translation or mostly rotation. Translational moves (approximately 1.0 m forward translation per second) were made when the robot was walking along a road in the fort, while rotational moves (approximately 22.5° yaw rotation per second) were made when the robot was at one of the four corners of rectangular path trying to change its moving direction. Since $\Delta t = 1$ s, the frame rate was 1 frame/s, which means the change of a view between two successive frames was approximately either a 1 m forward translation per frame or a 22.5° yaw rotation per frame. We placed a simulated trinocular stereo camera rig with 10cm baselines and a 70° field of view on the robot. To make the dataset somewhat challenging, we simulated the effects of a traveling on an imperfect outdoor surface, so that the robot's vertical position varied approximately ±0.04 m from 0, its pitch and roll varied ±2.5° from 0, and its yaw varied ±3° from its expected course.

The camera rig captured a trinocular image set at each time i, resuling in totally 321 image sets over the traveling. The images are gray scale images with 640 × 480 pixel in resolution. Images acquired by the simulated camera rig have no distortion. Hence no rectification was required as opposed to trinocular images obtained by real cameras in a real world experiment. Fig. 4.2 shows an image set captured in the simulation environment.

4.2 ST-SLAM in a Real Indoor Environment

In this experiment, we attempt to measure the robusness of the ST-SLAM system in the real world with various noise. We have adapted the algorithm for a real hardware system (a trinocular stereo head supported by a tripod) and tested it in a real indoor environment with image noise, camera calibration error, and odometry error.

We chose, as an indoor testbed, the Image and Vision Computing Laboratory at SIIT. The room is a typical laboratory with desks, bookshelves and computers. The shape



(a) (b) (c)

Figure 4.2: A sample trinocular image set captured in the simulated environment. (a) Reference image. (b) Horizontally aligned image. (c) Vertically aligned image.

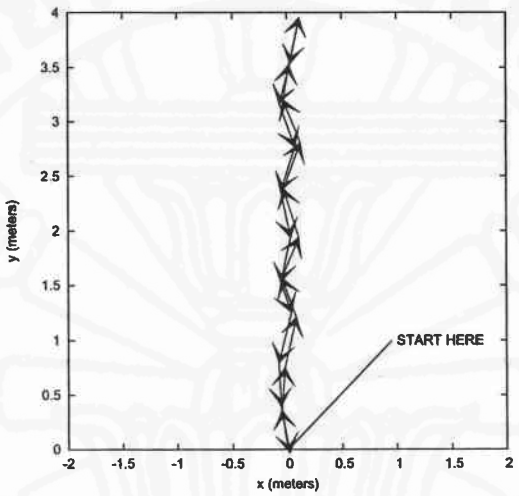


Figure 4.3: Ground truth path of the rig. The rig was initially placed at the origin. The path was composed of approximately 4 m of forward translation followed by a 180°-clockwise rotation and 4 m of forward translation.

of the room is approximately a rectangular parallelepiped 6.0 m in length, 8.5 m in breadth, and 3.0 m in height.

In the experiment, we simulated robot motions by manually moving the camera rig on a tripod and carefully measuring ground truth. The true path $\boldsymbol{s}_{0:t}$ was composed of approximately 4 m of forward translation over 10 moves, a 180°-clockwise rotation over 8 moves, and 4 m of forward translation over 10 moves, as shown in Fig. 4.3. The rig's pose $\boldsymbol{s}_t$ in world-coordinates at time t has six degrees of freedom: $\boldsymbol{s}_t = [s_{x,t}, s_{y,t}, s_{z,t}, s_{\phi,t}, s_{\theta,t}, s_{\psi,t}]^T$. Here the $s_{x,t}$ and $s_{y,t}$ axes span a plane parallel to the floor of the lab, and $s_{z,t}$ is the vertical distance of the reference camera's origin from the ground plane. As part of ground truth data, we calculated the robot-relative true moves $\Delta\boldsymbol{s}_{1:t}$ from measured $\boldsymbol{s}_{0:t}$.

We manually added Gaussian noise to each measured ground truth motion $\Delta\boldsymbol{s}_i$ in $\Delta\boldsymbol{s}_{1:t}$, then took the result as the measured odometry motion $\boldsymbol{u}_{1:t}$. The error added is shown in Table 4.1. We used a multiplier α to adjust the degree of yaw error. The most realistic values of α for a typical robot would be between 1.0 and 3.0. We attempted to determine the performance of the method especially with respect to yaw error,

Table 4.1 Error Added to Ground Truth (std.)

	x	y	z	ϕ	θ	ψ
Translational (step 1–10 and 19–28)	4.0cm	4.0cm	2.5cm	1.0°	1.0°	2α°
Rotational (step 11–18)	1.0cm	1.0cm	0.5cm	0.5°	0.5°	4α°

[1] $\alpha = 0.2, 0.5, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0$

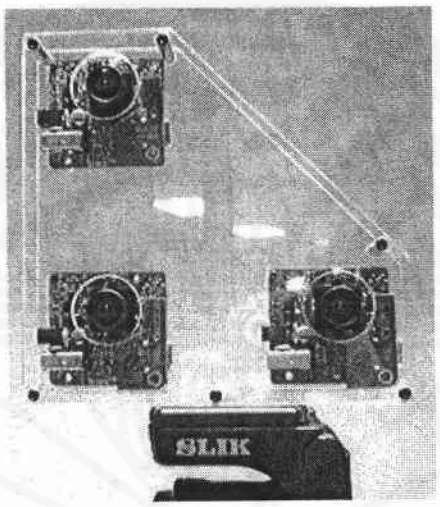


Figure 4.4 10cm-baseline trinocular camera rig used in the experiment.

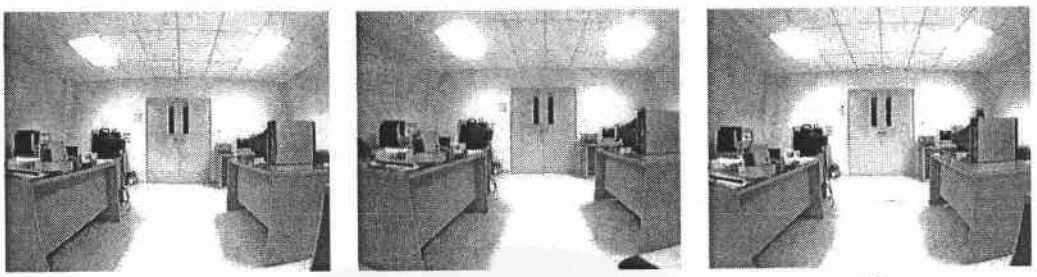
since in our case, good estimation of yaw is of critical importance to achieve successful localization and mapping. We fixed the noise level for the other components of s_t to reasonable levels depending on whether the move was translational or rotational.

The 10 cm-baseline trinocular stereo rig (shown in Fig. 4.4) supported by the tripod was 1.0 m high above the floor with pitch and roll equal to 0. We captured one trinocular image set at the initial position and immediately after each of the tripod's 28 moves. The effective view shift per frame was approximately 40 cm of forward translation or approximately 22.5° of yaw. Fig. 4.5 shows an image set captured in the lab with the trinocular camera rig.

4.3 ST-SLAM and Assumptions

Using the collected image sets and measured motions $u_{1:t}$ with different error levels as input, we ran ST-SLAM algorithm with 1, 10, 100, 1,000 and 10,000 particles. In order to compare our system's performance against the baseline, we also ran the same algorithm with one particle using pure odometry (odometry-only SLAM) and true moves (true-move-only SLAM) as the estimate of the tripod's motion. We quantitatively compared the results based on the likelihood of landmark observations and robot move errors. As a more quantitative evaluation, we visually inspected the reconstructed sparse 3D point landmark maps obtained in experiments.

Due to the flat floor or ground in both experiments, the height $s_{z,t}$, pitch $s_{\phi,t}$, and



(a) (b) (c)

Figure 4.5: A sample image set captured in the laboratory. (a) Reference image. (b) Horizontally aligned image. (c) Vertically aligned image.

roll $s_{\theta,t}$ were approximately equal to 0 throughout the experiments. So we assumed this as a priori knowlege on the motion model. Accordingly, in the particle sampling phase of RBPF, we filtered out particles whose currrent estimate of $s_{z,t}$, $s_{\phi,t}$, and $s_{\theta,t}$ significantly deviated from 0.

4.4 Log Likelihood as a Performance Measure

At each time step t, we sample the next move from the proposal distribution for each particle. Based on the sampled move and current 3D landmark observations, we calculate the likelihood of the new observation by comparing previously observed landmarks and the newly observed landmarks. Using more particles means we have a higher probability of sampling moves that are close to the true move. Therefore we generally expect that the likelihood (and its logarithm) should improve as we increase the number of particles.

In order to compare results across different SLAM configurations, we calculated the sum of the log likelihood over the entire move steps for each particle. This sum, which we call the *accumulated log-likelihood*, is a good measure conveying how consistent the observed landmarks are with the estimated moves. So a larger accumulated log likelihood means that we achieved the better localization and mapping. For each SLAM experiment, we chose the particle that had the best accumulated log likelihood.

4.5 Move Error as a Performance Measure

Log likelihood is a performance measure based on the cosistency of observed landmarks and the estimated moves. However, observations are dependent on the sensor model that is generally imperfect due to linearization and noise. To conduct a more direct and precise comparison of localization performance, we introduce move errors. The idea is, for each move, to calculate the error of the estimated move and the corresponding true move (move error), and take its statistics over the all move steps. Since the ground truth moves are available in both simulation and real-world experiments, we can use move errors to compare the performances of ST-SLAM with different setups.

4.6 Data Evaluation Method

4.6.1 Hypothesis Testing for Statistical Significance

We compare move errors of odometry-only SLAM and ST-SLAM with various noise levels. This performance evaluation requires rigorous comparison based on standard statistical techniques. For each step $i \in 1, \ldots, n$ of robot motion, where n is the total number of steps in an experiment, we pair odometry-only move error e_i^o and ST-SLAM move error e_i^s by calculating the difference

$$d_i = e_i^o - e_i^s. \tag{4.1}$$

We assume that each d_i arises from the underlining normal distribution $\mathcal{N}(\mu; \sigma^2)$ with unknown mean μ and variance σ^2, where we denote the corresponding random variable as D, i.e.

$$D \sim \mathcal{N}(\mu; \sigma^2). \tag{4.2}$$

We compare the performance of odometry-only SLAM and ST-SLAM by applying two-tailed significance tests with the 0.05 type I error level to the samples of D obtained by (4.1). We let the null hypothesis H_0 and the alternative hypothesis H_a be

$$H_0 : \mu = 0, \tag{4.3}$$

$$H_a : \mu \neq 0. \tag{4.4}$$

When the test result is that we fail to reject H_0, we conclude that ST-SLAM has no statistically significant difference in performance over odometry-only SLAM. When the test reveals that we reject H_0 in favour of H_a, we first check the sign of the mean of the samples $\bar{D} = \sum_{i=1}^{n} d_i/n$, then we conclude that ST-SLAM performs better than odometry-only SLAM if the sign is positive, or we conclude that SLAM perform worse than odometry-only SLAM if the sign is negative.

The choice of significance tests is based on n, the number of samples available from experiments. In our cases, $n = 320$ in the simulation outdoor experiment, and $n = 28$ in the real indoor experiment. We use the Z-test for the former and the T-test for the latter. In the following sections, we explain the two types of significance tests.

4.6.2 Z-test

The Z-test is a test of statistical significance used to determine whether the mean of samples is significantly different from the mean of the underlining population of the samples. The population is assumed to be a normal distribution with a known variance, or if the variance is unknown, the size of the samples should be large enough for the variance of the population to be well approximated by the variance of the samples.

Let X_i, $i \in 1, \ldots, n$ be n random samples that arise from the common normal distribution $\mathcal{N}(\mu; \sigma^2)$ with mean μ and known variance σ^2. We then define the *standardized test statistic* Z as

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \tag{4.5}$$

where $\bar{X}$ is the mean of the random samples: $\bar{X} = \sum_{i=1}^{n} X_i/n$. Due to the random Gaussian variables X_i, the distribution of $\bar{X}$ is $\mathcal{N}(\mu; \sigma^2/n)$ and the distribution of Z is $\mathcal{N}(0;1)$ [31].

We define the null hypothesis H_0 and the alternative hypothesis H_a as (4.3) and (4.4). We then define a two-tailed test with significance level α as

- Reject H_0 in favor of H_a if $Z < -z(\alpha/2)$ or $Z > z(\alpha/2)$,
- Accept H_0 otherwise.

$z(\alpha/2)$ is the *upper* $100(\alpha/2)$ *percentage point* of the $\mathcal{N}(0;1)$ distribution. In actual tests, we calculate Z based on (4.5), where $\bar{X}$ is obtained from data, $\mu = 0$ is due to the null hypothesis, σ is estimated from data, and n is known in the experiment.

4.6.3 T-test

The T-test is a test of statistical significance used to determine whether the mean of a small number of samples is significantly different from the mean of the underlining normal population with unknown variance. Unlike the Z-test, due to the small number of samples (typically, less than 30), we cannot approximate the variance of the population from data. However the ratio

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \tag{4.6}$$

where $S^2 = [1/(n-1)] \sum (X_i - \bar{X})^2$ is an unbiased estimator of variance σ^2, is known to have a well-defined distribution that is independent of σ [31]. This distribution is called the *Student's t-distribution with* $r = n - 1$ *degrees of freedom*, explicitly written as

$$h(t) = \frac{c}{(1 + t^2/r)^{(r+1)/2}}, \tag{4.7}$$

where the normalizing constant c is given by

$$c = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\,\Gamma\left(\frac{r}{2}\right)}. \tag{4.8}$$

The ratio T is referred to as the T *test statistic*. Using this test statistic and the same hypotheses H_0 and H_a introduced in the Z-test, we define a two-tailed test with significance level α as

- Reject H_0 in favor of H_a if $T < -t(\alpha/2)$ or $T > t(\alpha/2)$,
- Accept H_0 otherwise,

where $t(\alpha/2)$ is the upper $100(\alpha/2)$ percentage point of the t-distribution with r degrees of freedom.

4.6.4 Bonferroni Correction

When we conduct multiple significance tests upon the same sample data set, the probability that we reject each null hypothesis of the tests by chance becomes larger, which leads to falsely giving significance to some hypothesis. To correct this, we use the *Bonferroni correction* method [32].

Applying the Bonferroni correction to our cases implies that we should use α/n type I error level for each individual significance test in order to achieve a family of n significance tests with the α type I error level to the same sample data